Synthetic Biology Approaches to Engineering Human Cells

Synthetic Biology Approaches to Engineering Human Cells

A dissertation presented

by

Jason Jakob Lohmueller

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Microbiology and Molecular Genetics

Harvard University

Cambridge, Massachusetts

January 2013

Dissertation Advisor: Professor Pamela A. Silver Jason Jakob Lohmueller

Synthetic Biology Approaches to Engineering Human Cells

Abstract

The field of synthetic biology seeks to revolutionize the scope and scale of what is currently feasible by genetic engineering. By focusing on engineering general signal processing platforms using modular genetic parts and devices rather than 'one-off' systems, synthetic biologists aim to enable plug-and-play genetic circuits readily adaptable to different contexts. For mammalian systems, the goal of synthetic biology is to create sophisticated research tools and gene therapies. While several isolated parts and devices exist for mammalian systems there are few signal processing platforms available. We addressed this need by creating a transcriptional regulatory framework using programmable zinc finger (ZF) and TALE transcription factors and a conceptual framework for logical T-cell receptor signaling.

We first engineered a large set of ZF activator and repressor transcription factors and response promoters. ZFs are scalable elements as they can be engineered to bind to given DNA sequences. We demonstrated that we could 'tune' the activity of the ZF transcription factors by fusing them to protein homo-dimerization domains and by modifying their response promoters. We also created OR and NOR logic gates using hybrid promoters and AND and NAND logic gates by reconstituting split zinc finger activators and repressors with split inteins.

Next, using a computational algorithm we designed a series of TALE transcriptional activators and repressors predicted to be orthogonal to all 2kb human

promoter regions and thus minimally interfere with endogenous gene expression. TALEs can be designed to bind to even longer DNA sequences than ZFs, however off-target binding is predicted to occur. We tested our computationally designed TALEs in human cells demonstrating that they activated their intended target genes, but not their likely endogenous off-target genes, nor synthetic promoters with binding site mismatches.

Finally, we created a conceptual framework for logical T-cell-mediated killing of target cells expressing combinations of surface antigens. The systems consist of conventional and novel chimeric antigen receptors (CARs) containing inhibitory or co-stimulatory cytoplasmic signaling domains. In co-incubation assays of engineered T-cells with target cells we demonstrated a functioning OR-Gate system and progress toward development of a functional NOT-Gate system using the CD300a and CD45 inhibitory receptor domains.

Table of Contents

ACKNOWLEDGMENTS

For the work I present in this disseration, I would like to thank:

My family: Mom, for always believing in me, inspiring me in everything that I do, and having blessed me with such a happy life. Elisabeth, for setting an example for what it means selflessly help and serve others and for always being there to talk to. Kirk for sparking my interest in biology at an early age, and for being a source of brotherly and scientific wisdom. Grampy for always making sure I was on time for my train for my summer rotation.

To Lisa: for bringing so much wonderful into my life in so many ways.

To my friends:

David for the science chats and hilarity.

Rex for always being a great friend.

Abhishek and Tom, my coffee room collaborators, for being great friends and colleagues.

To Yvonne for teaching me everything I know about immunology wet-lab work and for being an amazing and understanding mentor and colleague.

To my dissertation advisory committee: Keith Joung, Steve Elledge, and Galit Lahav for your invaluable advice and comments over the years.

To my dissertation exam committee: Keith Joung, Richard Mulligan, Jagesh Shah and Wilson Wong, for your reading of this work.

To David Cardozo: for being the voice of reason and helping me through a tough time.

To Kathy: for always making everything in lab go so smoothly, always.

To the BBS office and Systems Biology Administrators, for helping with so many things along the way.

To everyone in the Silver lab, especially Keiji, Florian, Tyler, Joe T, it's been an amazing ride!

Finally to Pam, for always pushing me further in my ambition and creativity, for understanding and support through a difficult time, and for running the best lab anywhere.

For Mom.

Chapter I

Introduction: transcription factor-based mammalian synthetic gene circuits: design strategies and outlook

ABSTRACT

Mammalian synthetic gene circuits have the potential to directly impact human health in the form of sophisticated new research tools and gene therapies. Transcription factor-based circuits represent some of the most common and best-studied mammalian gene circuits, and recent technological advances have made their future appear even more promising. Much progress has been made toward expanding the repertoire of biomolecules that transcriptional circuits can interrogate. Additionally, several different types of computations have now been implemented with these circuits, including: logic devices, feedback loops, half-adders, and switches. In creating these circuits many key design strategies and considerations have emerged, offering genetic engineers a veritable toolkit for circuit generation. While thus far the scale and predictability of mammalian synthetic gene networks has been limited, the recent development of computational frameworks and targeted genome engineering tools will help to overcome some of these challenges. This chapter reviews current approaches to designing transcription factor-based mammalian synthetic gene circuits and discusses the future outlook of these circuits while highlighting mammalian-specific engineering challenges and potential solutions.

A promising outlook for mammalian synthetic gene networks

Already established as indispensible research tools and beginning to enter the clinic, mammalian synthetic gene networks are now poised to play an even greater role in human health. Researchers created the first mammalian synthetic gene circuit over 25 years ago when they demonstrated that the Lac repressor (LacI) from lambda phage could control transgene expression in mammalian cells in an IPTG-responsive manner (Brown, Figge et al. 1987). Since that time this LacI system and similar tramscription factor (TF) inducible gene expression systems, most namely the tetracycline/doxycycline inducible systems, have revolutionized studies of mammalian gene function, granting control over the timing and expression level of transgenes (Gossen and Bujard 1992; Urlinger, Baron et al. 2000). The scope of what circuits can do has also expanded far beyond these inducible systems. Circuits can now interrogate and respond to a vast array of biomolecular inputs, including: promoter-specific transcription, miRNAs, intracellular and surface proteins, light, post-translational modifications, and various chemical inducers (Culler, Hoff et al. 2010; Nissim and Bar-Ziv 2010; Toettcher, Voigt et al. 2011; Xie, Wroblewska et al. 2011; Leisner, Bleris et al. 2012). Researchers have developed several different approaches to integrate and process these signals. Using various genetic computing elements such as TFs, catalytic RNAs, and post-translational effector proteins, researchers have created a variety of circuit architectures including logic networks, feedback loops, and genetic switches (Deans, Cantor et al. 2007; Rinaudo, Bleris et al. 2007; Chen, Jensen et al. 2010; Leisner, Bleris et al. 2010; Burrill, Inniss et al. 2012). The recent development of general computing frameworks, large sets of mutually orthogonal genetic regulatory components, will likely lead to the development of higher-

order gene circuits that will grant even greater specificity and timing of gene expression (Khalil, Lu et al. 2012; Leisner, Bleris et al. 2012; Lohmueller, Armel et al. 2012).

Researchers are now using many of these genetic components to create applied circuits with clinical significance, most notably in immune cells. In one recent study by Chen et al. RNA-based controllers were used to accomplish small molecule-regulated control over engineered T-cell proliferation in mice (Chen, Jensen et al. 2010). The use of RNA-controllers in the study was important, as unlike proteins RNAs don't undergo antigen presentation. In another recent study, Wei and Wong used bacterial effector proteins to tune T-cell activation and to create a synthetic pause switch to temporarily inhibit T-cell activation. These circuits could serve as safety switches for immune therapies (Wei, Wong et al. 2012). In yet another recent study, Kloss et al. constructed an AND gate using chimeric antigen receptors (CARs) to lyse target cells that have a combination of two antigens on their surface in mice (Kloss, Condomines et al. 2012). While single CAR systems have been in clinical trials for some time, this new system could be used to greatly enhance the specificity of CAR-based T-cell therapies (Porter, Levine et al. 2011). Other notable, non-immunological systems include circuits by Ye et al. and Kemmer et al. in which TF and receptor-based circuits were used to sense and prevent the onset of the metabolic diseases of hyperuricemia and metabolic syndrome, respectively, in animal models (Kemmer, Gitzinger et al. 2010; Ye, Daoud-El Baba et al. 2011; Ye, Charpin-El Hamri et al. 2013). Looking forward, there are also many challenges to address before realizing the full potential of mammalian synthetic gene networks. In general mammalian systems are inherently complex making the engineering

process slower and circuit behaviors more difficult to predict. Some of these challenges are beginning to be addressed while others are still awaiting solutions.

In this chapter we review current approaches to generating mammalian transcription factor-based synthetic gene circuits. We focus on general design principles including the types of inputs, design of regulatory components, and different circuit architectures while discussing important circuit parameters. Finally, we contemplate the future of mammalian synthetic gene circuits including potential future approaches and and mammalian-specific engineering challenges and potential solutions.

TRANSCRIPTION-BASED GENE CIRCUITS

Transcriptional Regulators

By far the largest number of mammalian synthetic circuits have been transcription-based systems. This is probably because transcription factors were the first synthetic regulators used in mammalian circuits and conceptually, transcription-based systems are intuitive and relatively simple to design and implement (Kramer, Fischer et al. 2004; Kramer, Fischer et al. 2005; Kramer and Fussenegger 2005). Transcriptional circuits consist primarily of well-defined defined positive and negative transcription factor (TF) regulators and their cognate regulatory promoters. TFs are most often comprised of DNA binding domains (DBDs) fused to transcriptional activation or repression domains for positive and negative regulators, respectively. The most commonly used activation domains are the VP16 activation domain from Herpes Simplex Virus TK activator, VP64 four tandem copies of VP16, and the activation domain from the p65 protein of the NFkB transcriptional activator (Hurt, Thibodeau et al. 2003). The

most common negative regulatory domain is the KRAB repression domain (Beerli and Barbas 2002; Haynes and Silver 2011). The underlying biochemical regulatory mechanisms of these domains have largely been worked out, however, several papers suggest roles for these domains in chromatin remodeling. The mechanisms for these functions and impact on circuit behavior have not been fully uncovered (Urrutia 2003; Haynes and Silver 2011).

The bulk of DBDs in the first synthetic regulators were derived from lower organisms such as bacteria and yeast and more recently from programmable DBDs such as zinc finger and TALE proteins. Of the early synthetic regulators LacI, TetR, and Gal4. LacI and TetR are unique as there are small molecule inhibitors IPTG, and doxycycline that can act to reduce their DNA binding activity activity. TetR was also modified through mutagenesis to create the reverse tetracycline transactivator (rtTA) such that its DNA binding activity is instead induced by doxycycline (Gossen and Bujard 1992). These factors have well defined DNA binding sequences that are used in generating response promoters. Other non-programmable synthetic TFs used in mammalian circuits include the streptogramin and macrolide inducible factors PIP, and ETR (Kramer, Viretta et al. 2004).

Recent TF regulator designs have focused on the construction and use of DBDs that are ‘programmable’ in the sense that researchers can design them to bind to different DNA sequences. The most prominent examples are zinc finger (ZF) and TALE proteins. ZF DBDs most often consist of arrays of 3-4 individual fingers that each bind to 3bp of DNA sequence (9-12bp DNA binding sequence in total) (Hurt, Thibodeau et al. 2003). While the original modular assembly hypothesis positing that fingers known to bind to 3-

peat sequences can be strung together to create larger binding proteins with combined specificity was discredited, this theory was recently re-written and validated including new rules taking into account context-dependent effects of positions of the fingers (Sander, Reyon et al. 2010; Sander, Dahlborg et al. 2011). Now three finger ZF proteins can be efficiently generated without the need for lengthy selection processes. Several positive and negative ZF TFs regulators have been generated and shown to function in mammalian cells. ZF TFs have been shown to be regulated by small molecules through the use of induced dimerization protein domains, and tunable by adding constitutive homo-dimerization domains of different strengths on the ZFs (Beerli, Schopfer et al. 2000; Lohmueller, Armel et al. 2012). As ZFs can only bind to 9-12bp sequences with specificity they will likely have off-target binding in the host cell's genome. On the upside, ZFs are relatively small proteins, and each 3-finger ZF is encoded by only ~300bp of DNA.

TALE proteins represent another class of programmable DBDs. TALEs were recently discovered to have a well-defined binding code in which the individual repeat domains that make up TALE DBDs each have a preference for binding to a single nucleotide with high specificity (Boch, Scholze et al. 2009; Moscou and Bogdanove 2009). By stringing together the repeat domains with known specificities, researchers can very easily design TALEs that can bind specifically to longer sequences. Researchers traditionally design TALEs to bind 18-20bp meaning they are capable of binding to DNA sequences twice as long as ZFs (Garg, Lohmueller et al. 2012). The caveat however, is that the TALE repeats do not have perfect specificity for a single nucleotide, and thus some off-target binding is expected to occur (Moscou and Bogdanove 2009). Both TALE

activators and repressors have been shown to work effectively in mammalian cells (Garg, Lohmueller et al. 2012). In one of our papers we used a computational approach to generate TALEs that are not expected to bind to any endogenous human promoter regions even after taking predicted off-target binding into account, creating an ideal set of regulators to use in circuits. As the research on TALE DBDs is still in its infancy, it is possible that large scale expression studies will be useful in determining the effects of these regulators on the cells' endogenous gene expression. Like most TFs the strength of TALE expression can be modulated by increasing the number of binding sites in the TALE promoter. To decrease activity, one can created mutated binding sites at different positions. One downside to using TALEs is that they are very large proteins to work with, and contain highly repetitive DNA sequences. To reduce this difficulty there have been several effective cloning schemes developed for TALE assembly (Morbitzer, Elsaesser et al. 2011; Reyon, Tsai et al. 2012; Schmid-Burgk, Schmidt et al. 2012). However, TALEs are still too long to work effectively in viral vectors with limited sizes or high sensitivity to repetitive DNA (Holkers, Maggio et al. 2012).

Response Promoters

To create activatable response promoters TF binding sequences are placed upstream of a minimal/core mammalian promoter. The most common minimal promoters are TATA boxes derived from the CMV promoter and HSV-TK viral promoters and or the CMVmin promoter which contains extra core promoter elements (Agha-Mohammadi, O'Malley et al. 2004). The strength of this minimal promoter is an important factor for regulating both the background transcriptional activity in the absence of a TF and the

final maximum activated transcriptional activity in the presence of a TF. The choice of the minimal promoter is an often-underappreciated design consideration. In an interesting study Juven-Gershon et al. created a high expressing 'super core' promoter by combining all known mammalian core promoter elements (Juven-Gershon, Cheng et al. 2006). These elements could perhaps be used to individually or in combination to create new minimal promoters with variable activities.

Repressible promoters consist of constitutive promoters containing TF binding sites somewhere in the promoter, most often at the transcriptional start sites or at known endogenous TF binding sites. The repressor TF acts by binding to the sequence in the promoter and inhibiting the transcription either through binding or through the activity of its repression domain (Beerli and Barbas 2002). The maximum activity of the repressible promoter is defined by the upstream constitutive promoter. The rules for the 'activity range' of repression domains such as the KRAB domain, where the TF needs to bind to inhibit transcription, have not been elucidated. Future studies using designer TFs, for which TFs could be designed to bind anywhere in a promoter, will likely uncover some of these characteristics. For both activatable and repressible promoters, the number of binding sites in the promoter is an important design variable that defines the response strength of the promoter to the TF (Lohmueller, Armel et al. 2012).

Inputs to TF Circuits

Inputs for transcriptional circuits are diverse and can in theory include anything capable of affecting the expression or activity of a TF regulator. That being said while linking an input to a TF regulator can be conceptually simple, in practice the generation

of a properly functioning circuit often requires precise tuning and an input with a wide dynamic range. Experimentally demonstrated inputs include: endogenous promoter activities, endogenous TF activities, miRNAs, small molecules, light, proteases and cell surface receptor ligands (Wehr, Laage et al. 2006; Xie, Wroblewska et al. 2011; Li, Moore et al. 2012). While some of these circuits have intermediary steps between the original signal and the final transcriptional output, we are considering anything that can eventually affect the TF regulator as a potential circuit input. Potential inputs that as of yet have not been demonstrated are also discussed.

Promoter activity can be used as an input by placing the promoter upstream of a TF-regulator to drive its expression. Endogenous gene expression data is a powerful classifier for many different biological states including cell type and various disease states. There are many promoters that have been discovered from such data and cloned and experimentally verified to regulate a transgene in a specific manner (Trinklein, Aldred et al. 2003; Kim, Barrera et al. 2005; Zhang, Markus et al. 2012). These are ideal promoters to use as inputs as they have already been isolated. Potential promoters from qPCR data or large expression sets such as microarrays or RNAseq can also be used, however there are several considerations in this case. First, the gene's RNA expression is the result of a promoter and other factors such as RNA stability, miRNAs targeting the transcript, and chromatin modifications. By only taking a gene's promoter, it is possible that some of this regulation will be lost. Including the UTRs or introns of the gene in the expression of the TF could help to recapitulate some of this regulation, however this approach has not yet been reported. It is also often difficult to define the full regulatory regions of the promoter. Enhancer elements containing endogenous TF sites can be

located up to 100kb away from the expressed gene, and it is difficult to define DNA elements necessary to recapitulate native chromatin structure of the gene (Zhang, Markus et al. 2012). In general it is better to err on the long side when designing promoter sequences, however this option is not always possible as DNA sequence size is often a major limitation when making a circuit.

Endogenous TFs have also been used as circuit inputs by creating synthetic promoters responsive to the endogenous TFs. These promoters are generally created like synthetic circuit TF promoters with the endogenous TF binding sites upstream of a minimal promoter (Li, Moore et al. 2012). Many endogenous TF binding sites have been experimentally discovered, however, TF binding is not always enough to induce or repress transcription. The regulation of genes by endogenous TFs is complex and TFs often regulate genes in combination with other endogenous factors. Thus, for some TFs it could be better to use endogenous promoters known to respond strongly to the TF in question (Whitfield, Wang et al. 2012).

Small inhibitory RNAs such as miRNAs can be interrogated as inputs by placing RNA recognition sites in the 3' UTR of the TF transcript. Perfectly matching sequences will regulate TF expression by degrading the transcript while slightly mismatched sequences will inhibit translation of the TF regulator from the transcript. These sites can be multimerized to increase the inhibitory effect of the small RNAs (Leisner, Bleris et al. 2010).

Small molecules and metabolites have also been interrogated as TF circuit inputs. These are most often interrogated by directly acting on a TF-regulator inducing or inhibiting its DNA binding or assembly of split-TF protein fragments. Interrogated small

molecules include IPTG, Doxycycline, 4HT, uric acid, rapamycin, macrolide, and streptogramin (Deans, Cantor et al. 2007; Kemmer, Gitzinger et al. 2010; Leisner, Bleris et al. 2010; Ye, Charpin-El Hamri et al. 2013). Light has also been used to trigger assembly of split-TF protein fragments to serve as an input in this way (Bacchus and Fussenegger 2012). Expanding the number of TFs inducible by metabolites is an active field of interest.

Finally, cell surface signaling can also serve as an input to a synthetic TF circuit. While this sensing must be indirect there are a couple of different methods developed for interrogating cell surface signaling. In some cases these methods have been demonstrated for a very specific set of regulators, but some could be adopted as general strategies for other receptors. In some of the most general methods, the TEV protease is split into two fragments with one fragment fused to the receptor along with a TEV-cleavage sequence and a TF. For G-Protein Coupled Receptors (GPCRs) the TEV is re-consituted using a recruitment method, in which beta-arrestin is fused to the other half of split TEV. Upon receptor activation the beta-arrestin TEV is recruited to the receptor leading to cleaving the TEV cleavage peptide and release of the tethered TF. Similar methods splitting TEV and fusing it to two receptors that dimerize upon activation has also been demonstrated. One can also indirectly measure the cell-surface activation by monitoring activity of endogenous TFs known to be activated by cell surface receptor signaling. This strategy can potentially lead to false positives as multiple cell signals can lead to TF activation, but there are often endogenous gene promoter sequences specific to certain cell-signaling pathways (Wehr, Laage et al. 2006; Barnea, Strapps et al. 2008).

Other potential inputs for a TF circuit that have not yet been used in TF circuits include endogenous nuclear proteins and small molecules/ metabolites that can bind to RNA molecules (Yen, Svendsen et al. 2004; Murphy, Mostoslavsky et al. 2006). In a recent study by Culler et al. transgenes were fitted with an exon containing a protein-binding aptamer sequence flanked on either side by two introns (Culler, Hoff et al. 2010). In the presence of the protein being sensed this protein binds to the aptamer and either leads to splicing or inhibition of splicing depending on the intron-exon-inton design. This system could be adapted to regulate a TF as part of a circuit.

TF computation

Transcriptional regulators have been shown to be modular and combining these regulators to generate logic gates, feedback loops, and larger-scale circuits is conceptually simple. After linking the desired inputs to the presence or activity of the TF regulators then the response promoters must be configured and tuned to perform the desired computation. The most common types of circuits built to date are logic circuits. These circuits allow the activation of an output gene or genes in response to the presence or absence of multiple inputs in a logical fashion. Two- input OR, NOR, AND, NAND, and A AND NOT B gates, and most recently half-adder circuits, have been demonstrated using TF-regulators (Kramer, Fischer et al. 2004; Auslander, Auslander et al. 2012; Lohmueller, Armel et al. 2012). OR logic gates are the simplest to design and have been generated by multiple groups by linking the presence of two inputs to two distinct TFs while fitting the activatable output promoter with binding sites for both TFs. It has also been generated by having 2 output promoters, each individually activated by the single

TFs (Benenson 2011). This design leads to a more analogue response in which the total output gene expression is higher when both TFs are present. A NOR gate can be created in a similar fashion to an OR gate by connecting inputs to two distinct repressor TFs. Creating AND and NAND logic is less conceptually straightforward, and cannot be accomplished simply by engineering response promoters, but there are a established methods available. The most common method is to use a 2-hybrid approach in which a TF is split into two proteins, most often into a DBD fragment and an activation domain fragment, and each protein is fused to a protein known to bind to the other protein. Upon expression of both proteins the two TF halves interact and form a complete factor turning on expression of the output gene. We have demonstrated another similar method using split inteins to splice together the TF halves into a complete TF factor. We also demonstrated creating a NAND gate using the split intein method, by instead splitting a repressor TF. Only when both fragments are present is the output protein repressed. The intein method has the advantage of performing computation with a complete TF factor, however this approach has the disadvantage of requiring use multiple intein-based AND gates in a cell, multiple, orthogonal split inteins would be required (instead of multiple protein-interaction domains in the case of the 2-hybrid approach.) In one study by different positive and negative TFs were used to integrate miRNA signals (Nissim and Bar-Ziv 2010; Lohmueller, Armel et al. 2012).

Another type of circuit that has been demonstrated is a feedback loop. Positive feedback loops can be used to amplify or sustain a response, whereas negative feedback loops can be used to generate pulses or pauses in output. A transcriptional positive feedback loop has been generated in mammalian cells and used as a ‘memory loop’

circuit. In this circuit a transient stimulus leads to the activation of an activator TF activates an output gene and more expression of itself (Burrill, Inniss et al. 2012). This system was used to effectively track cells that had experienced hypoxia and mutagen in culture. One downside to the circuit compared to other memory circuits such as recombinase circuits is that the stable signal is eventually lost over time. However these systems have the advantage of being tunable and reversible. While a TF-based negative feedback loop has not yet been shown in mammalian cells, one such circuit has been demonstrated in bacterial cells and the design could be transferrable to mammalian cells (Basu, Mehreja et al. 2004).

Many switches have also been developed using TF circuits. In the simplest case inducible TFs such as LacI and TetR can be considered switches as they have low leakiness and can be strongly induced. A tight repression switch was also created by Deans et al. combining TF repression with RNAi greatly minimizing leakiness in the Off-state (Deans, Cantor et al. 2007). We also demonstrated that this double regulation is possible using a dual expression miRNA and TALE module that would allow for tight regulation of any transgene or endogenous gene (Garg, Lohmueller et al. 2012). The positive feedback loop can also be thought of a stable switch as stable expression results from a transient stimulus. A bi-stable toggle switch has also been demonstrated using TF repressors. This system consists of two repressors, each inhibiting the activity of the other repressor. When one stimulus is given it induces one of the repressors leading repression of the other factor stably locking the system into that expression state. When the other stimulus is given the currently active TF repressor is inhibited, allowing the 2nd TF to be expressed and simultaneously repress the first TF similarly maintain the expression state.

Challenges for TF Circuits

While much progress has been made constructing transcription-based mammalian synthetic gene networks, many challenges remain. Recent technological advances will address some of these challenges while others will likely require new approaches. Many of these challenges are universal to the engineering of synthetic gene networks across all organisms, however there are also several challenges specific to mammalian circuits. The mammalian specific challenges are largely the result of the higher complexity of mammalian cells compared to lower organisms.

New general genetic engineering approaches from outside the field of synthetic circuits will likely address many current technical problems. The process of testing systems stably in mammalian cells is inherently difficult and lengthy as they do not generally maintain stable episomal DNA and mammalian genome engineering is difficult. Advances to recombinant DNA cloning such as the various isothermal DNA assembly methods are already helping to speed up the circuit generation process (Gibson, Young et al. 2009). Additionally, advances to mammalian cell genome engineering including zinc finger and TALE nucleases methods and the even more recent Cas9 system should greatly improve the ability to make targeted genome modifications (Miller, Tan et al. 2011; Sander, Dahlborg et al. 2011; Mali, Yang et al. 2013). These methods will be integral for enabling clinical-grade therapies in which genes and circuits can be integrated in safe harbor loci, ensuring that they interfere minimally with endogenous gene regulation. Researchers working on mammalian circuits should also take advantage of more stable cell line creation methods such as lentivirus, adenovirus, and the various transposon-mediated integration methods available (Bakota, Brandt et al. 2012; Di

Matteo, Matrai et al. 2012). Circuit copy number is currently an under-investigated system variable, especially given the major impact of human gene copy number on endogenous gene regulation.

In lieu of stable integration, currently many circuits and circuit components are tested or demonstrated entirely using transient transfections of DNA plasmids. While these methods offer quick system readouts, generally 24-48hrs post-transfection, in general transient circuit behaviors aren't always representative of a stable integrated system. This result is largely due to the issue that cells receive different numbers of plasmids, and in the case of multi-plasmid transfections cells will not always receive all plasmids being transfected. Additionally, transfection can be inefficient for some cell lines and primary cell types. There are potentially some scenarios where transient transfections are the final desired product, such as in the case of a transient therapeutic like something involving liposome-delivered DNA plasmids (Kim and Eberwine 2010). Additionally, there are efforts to predict more information from transients using advanced mathematical modeling. However in general stable cell lines are desirable and plasmid transfections will shift to being used more as first-pass check for qualitative component or system performance.

The issue of cell type heterogeneity is another major mammalian-specific problem. In general most mammalian systems networks have been demonstrated in an easy to work with cell line such as HEK293 cells or HELA cells. While the assumption is that circuits will perform similarly in different cell types this assumption is of course not always true. This issue is especially prominent when creating circuits to interrogate endogenous inputs. Once again genome engineering methods will help to allow

researchers to create circuits in more cell types including primary cells. Hopefully researchers will also be able to move beyond proof-of-principle circuit generation and instead build off of existing computing frameworks to create applied systems in desired cell types.

The issue of scale is faced when creating synthetic circuits across all organisms, Currently, few mammalian circuits with more than two inputs have been constructed, and circuits that can perform complex functions such as counting have not yet been demonstrated. The small circuit size is the result of many factors including the genome engineering challenges already mentioned. The issues of interacting circuit elements, through gene read-through, chromatin silencing, and chromosomal position effects are also important hindrances for multi-component circuits. Of note the largest scale circuits to date have been demonstrated using transient methods which in addition to being quick to test are not subject these challenges (Auslander, Auslander et al. 2012). It is well documented that neighboring transgenes on plasmids or integrated near each other in a chromosome tend to interfere through transcriptional read through or anti-sense mechanisms. There are a few safe-harbor loci known for getting rid of endogenous gene effects on mammalian circuits (Hermann, Maeder et al. 2012). However this doesn't account for circuit element interference, and it would be way more time efficient to insert entire circuits into a single locus. There have been insulator sequences reported to reduce this interference, however, these elements have not been widely used and they appear to have general silencing effects (Walters, Fiering et al. 1999). The selection of synthetic sequences or deeper analysis of mammalian insulator elements would greatly help this scaling process for which there is currently no clear-cut solution. There is also the issue

of how many circuit parts cells can handle without inducing cellular toxicity. Most circuits currently use high-expressing promoters as circuit components whereas endogenous gene regulation is often more subtle. It is very possible that multiple strong promoters in a cell will always test the limits of transcription and translation and prove to be toxic. It is ideal therefore to create circuits that will generate circuit outputs more in accordance with endogenous regulation. For instance it would be desirable in some cases to stably down-regulate genes using induced heterochromatin rather than constant high expression of an inhibitory TF. Currently the largest scale circuits use regulatory components that act at different levels of circuit activity, rather than relying solely on TFs (Auslander, Auslander et al. 2012). It is possible that as diverse regulator approach will be the prevailing approach to generating large-scale circuits in the future.

Of yet, no large-scale screening of circuit configurations have been demonstrated on a circuit-level scale for mammalian circuits, however, as creating complex circuit architectures is attempted, and finer tuning of circuit elements is required, it is likely that screening will become an important element of mammalian circuit generation. Retroviral and lentiviral mammalian libraries mammalian gene and shRNA libraries have been demonstrated, and thus the technical capability to create circuit libraries exists (Silva, Li et al. 2005). Additionally, the framework generation of many components with variable activities provides elements to use in such screening.

Circuits aiming for the clinic face yet another set of safety and efficacy challenges. One major safety challenge includes minimizing the effect of the circuit elements on endogenous gene expression, for example so as to not up-regulate oncogenic gene expression. Genome engineering methods and orthogonal factors like the orthogonal

TALEs that we demonstrated will be helpful toward creating insulated circuits with minimal effects on gene regulation. There is also the overlooked challenge of immune tolerance of non-self circuit components by a patient's immune system. Most circuits are more immediately useful as research tools and so this isn't a major concern, however as the field advances and more circuits intended for the clinic, it will be a challenge. It is likely that a patient's immune system would be activated to kill off the cells expressing circuit components. The immune system is likely to be more tolerant to RNA-based gene circuits, researchers working on circuits with proteins don't yet take this into account. It could be possible to engineer circuit components that are not recognized by the immune system. The new focus on immune circuits should help to start discussions in the synthetic biology community about clinical-ready circuits to overcome these challenges.

REFERENCES

Agha-Mohammadi, S., M. O'Malley, et al. (2004). "Second-generation tetracycline-regulatable promoter: repositioned tet operator elements optimize transactivator synergy while shorter minimal promoter offers tight basal leakiness." The journal of gene medicine **6**(7): 817-828.

Auslander, S., D. Auslander, et al. (2012). "Programmable single-cell mammalian biocomputers." Nature **487**(7405): 123-127.

Bacchus, W. and M. Fussenegger (2012). "The use of light for engineered control and reprogramming of cellular functions." Current opinion in biotechnology **23**(5): 695-702.

Bakota, L., R. Brandt, et al. (2012). "Triple mammalian/yeast/bacterial shuttle vectors for single and combined Lentivirus- and Sindbis virus-mediated infections of neurons." Molecular genetics and genomics : MGG **287**(4): 313-324.

Barnea, G., W. Strapps, et al. (2008). "The genetic design of signaling cascades to record receptor activation." Proceedings of the National Academy of Sciences of the United States of America **105**(1): 64-69.

Basu, S., R. Mehreja, et al. (2004). "Spatiotemporal control of gene expression with pulse-generating networks." Proceedings of the National Academy of Sciences of the United States of America **101**(17): 6355-6360.

Beerli, R. R. and C. F. Barbas, 3rd (2002). "Engineering polydactyl zinc-finger transcription factors." Nature biotechnology **20**(2): 135-141.

Beerli, R. R., U. Schopfer, et al. (2000). "Chemically regulated zinc finger transcription factors." The Journal of biological chemistry **275**(42): 32617-32627.

Benenson, Y. (2011). "Engineering RNAi circuits." Methods in enzymology **497**: 187-205.

Boch, J., H. Scholze, et al. (2009). "Breaking the code of DNA binding specificity of TAL-type III effectors." Science **326**(5959): 1509-1512.

Brown, M., J. Figge, et al. (1987). "lac repressor can regulate expression from a hybrid SV40 early promoter containing a lac operator in animal cells." Cell **49**(5): 603-612.

Burrill, D. R., M. C. Inniss, et al. (2012). "Synthetic memory circuits for tracking human cell fate." Genes & development **26**(13): 1486-1497.

Chen, Y. Y., M. C. Jensen, et al. (2010). "Genetic control of mammalian T-cell proliferation with synthetic RNA regulatory systems." Proceedings of the National Academy of Sciences of the United States of America **107**(19): 8531-8536.

Culler, S. J., K. G. Hoff, et al. (2010). "Reprogramming cellular behavior with RNA controllers responsive to endogenous proteins." Science **330**(6008): 1251-1255.

Deans, T. L., C. R. Cantor, et al. (2007). "A tunable genetic switch based on RNAi and repressor proteins for regulating gene expression in mammalian cells." Cell **130**(2): 363-372.

Di Matteo, M., J. Matrai, et al. (2012). "PiggyBac toolbox." Methods in molecular biology **859**: 241-254.

Garg, A., J. J. Lohmueller, et al. (2012). "Engineering synthetic TAL effectors with orthogonal target sites." Nucleic acids research **40**(15): 7584-7595.

Gibson, D. G., L. Young, et al. (2009). "Enzymatic assembly of DNA molecules up to several hundred kilobases." Nature methods **6**(5): 343-345.

Gossen, M. and H. Bujard (1992). "Tight control of gene expression in mammalian cells by tetracycline-responsive promoters." Proceedings of the National Academy of Sciences of the United States of America **89**(12): 5547-5551.

Haynes, K. A. and P. A. Silver (2011). "Synthetic reversal of epigenetic silencing." The Journal of biological chemistry **286**(31): 27176-27182.

Hermann, M., M. L. Maeder, et al. (2012). "Evaluation of OPEN zinc finger nucleases

for direct gene targeting of the ROSA26 locus in mouse embryos." PloS one **7**(9): e41796.

Holkers, M., I. Maggio, et al. (2012). "Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells." Nucleic acids research.

Hurt, J. A., S. A. Thibodeau, et al. (2003). "Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection." Proceedings of the National Academy of Sciences of the United States of America **100**(21): 12271-12276.

Juven-Gershon, T., S. Cheng, et al. (2006). "Rational design of a super core promoter that enhances gene expression." Nature methods **3**(11): 917-922.

Kemmer, C., M. Gitzinger, et al. (2010). "Self-sufficient control of urate homeostasis in mice by a synthetic circuit." Nature biotechnology **28**(4): 355-360.

Khalil, A. S., T. K. Lu, et al. (2012). "A synthetic biology framework for programming eukaryotic transcription functions." Cell **150**(3): 647-658.

Kim, T. H., L. O. Barrera, et al. (2005). "Direct isolation and identification of promoters in the human genome." Genome research **15**(6): 830-839.

Kim, T. K. and J. H. Eberwine (2010). "Mammalian cell transfection: the present and the future." Analytical and bioanalytical chemistry **397**(8): 3173-3178.

Kloss, C. C., M. Condomines, et al. (2012). "Combinatorial antigen recognition with balanced signaling promotes selective tumor eradication by engineered T cells." Nature biotechnology **31**(1): 71-75.

Kramer, B. P., C. Fischer, et al. (2004). "BioLogic gates enable logical transcription control in mammalian cells." Biotechnology and bioengineering **87**(4): 478-484.

Kramer, B. P., M. Fischer, et al. (2005). "Semi-synthetic mammalian gene regulatory networks." Metabolic engineering **7**(4): 241-250.

Kramer, B. P. and M. Fussenegger (2005). "Transgene control engineering in mammalian cells." Methods in molecular biology **308**: 123-143.

Kramer, B. P., A. U. Viretta, et al. (2004). "An engineered epigenetic transgene switch in mammalian cells." Nature biotechnology **22**(7): 867-870.

Leisner, M., L. Bleris, et al. (2010). "Rationally designed logic integration of regulatory signals in mammalian cells." Nature nanotechnology **5**(9): 666-670.

Leisner, M., L. Bleris, et al. (2012). "MicroRNA circuits for transcriptional logic." Methods in molecular biology **813**: 169-186.

Li, Y., R. Moore, et al. (2012). "Transcription activator-like effector hybrids for conditional control and rewiring of chromosomal transgene expression." Scientific reports **2**: 897.

Lohmueller, J. J., T. Z. Armel, et al. (2012). "A tunable zinc finger-based framework for Boolean logic computation in mammalian cells." Nucleic acids research **40**(11): 5180-5187.

Mali, P., L. Yang, et al. (2013). "RNA-Guided Human Genome Engineering via Cas9." Science.

Miller, J. C., S. Tan, et al. (2011). "A TALE nuclease architecture for efficient genome editing." Nature biotechnology **29**(2): 143-148.

Morbitzer, R., J. Elsaesser, et al. (2011). "Assembly of custom TALE-type DNA binding domains by modular cloning." Nucleic acids research **39**(13): 5790-5799.

Moscou, M. J. and A. J. Bogdanove (2009). "A simple cipher governs DNA recognition by TAL effectors." Science **326**(5959): 1501.

Murphy, G. J., G. Mostoslavsky, et al. (2006). "Exogenous control of mammalian gene expression via modulation of translational termination." Nature medicine **12**(9): 1093-1099.

Nissim, L. and R. H. Bar-Ziv (2010). "A tunable dual-promoter integrator for targeting of cancer cells." Molecular systems biology **6**: 444.

Porter, D. L., B. L. Levine, et al. (2011). "Chimeric antigen receptor-modified T cells in chronic lymphoid leukemia." The New England journal of medicine **365**(8): 725-733.

Reyon, D., S. Q. Tsai, et al. (2012). "FLASH assembly of TALENs for high-throughput genome editing." Nature biotechnology **30**(5): 460-465.

Rinaudo, K., L. Bleris, et al. (2007). "A universal RNAi-based logic evaluator that operates in mammalian cells." Nature biotechnology **25**(7): 795-801.

Sander, J. D., E. J. Dahlborg, et al. (2011). "Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA)." Nature methods **8**(1): 67-69.

Sander, J. D., D. Reyon, et al. (2010). "Predicting success of oligomerized pool engineering (OPEN) for zinc finger target site sequences." BMC bioinformatics **11**: 543.

Schmid-Burgk, J. L., T. Schmidt, et al. (2012). "A ligation-independent cloning technique for high-throughput assembly of transcription activator-like effector genes." Nature biotechnology **31**(1): 76-81.

Silva, J. M., M. Z. Li, et al. (2005). "Second-generation shRNA libraries covering the mouse and human genomes." Nature genetics **37**(11): 1281-1288.

Toettcher, J. E., C. A. Voigt, et al. (2011). "The promise of optogenetics in cell biology: interrogating molecular circuits in space and time." Nature methods **8**(1): 35-38.

Trinklein, N. D., S. J. Aldred, et al. (2003). "Identification and functional analysis of human transcriptional promoters." Genome research **13**(2): 308-312.

Urlinger, S., U. Baron, et al. (2000). "Exploring the sequence space for tetracycline-dependent transcriptional activators: novel mutations yield expanded range and sensitivity." Proceedings of the National Academy of Sciences of the United States of America **97**(14): 7963-7968.

Urrutia, R. (2003). "KRAB-containing zinc-finger repressor proteins." Genome biology **4**(10): 231.

Walters, M. C., S. Fiering, et al. (1999). "The chicken beta-globin 5'HS4 boundary element blocks enhancer-mediated suppression of silencing." Molecular and cellular biology **19**(5): 3714-3726.

Wehr, M. C., R. Laage, et al. (2006). "Monitoring regulated protein-protein interactions using split TEV." Nature methods **3**(12): 985-993.

Wei, P., W. W. Wong, et al. (2012). "Bacterial virulence proteins as tools to rewire kinase pathways in yeast and immune cells." Nature **488**(7411): 384-388.

Whitfield, T. W., J. Wang, et al. (2012). "Functional analysis of transcription factor binding sites in human promoters." Genome biology **13**(9): R50.

Xie, Z., L. Wroblewska, et al. (2011). "Multi-input RNAi-based logic circuit for identification of specific cancer cells." Science **333**(6047): 1307-1311.

Ye, H., G. Charpin-El Hamri, et al. (2013). "Pharmaceutically controlled designer

circuit for the treatment of the metabolic syndrome." Proceedings of the National Academy of Sciences of the United States of America **110**(1): 141-146.

Ye, H., M. Daoud-El Baba, et al. (2011). "A synthetic optogenetic transcription device enhances blood-glucose homeostasis in mice." Science **332**(6037): 1565-1568.

Yen, L., J. Svendsen, et al. (2004). "Exogenous control of mammalian gene expression through modulation of RNA self-cleavage." Nature **431**(7007): 471-476.

Zhang, J., J. Markus, et al. (2012). "Three murine leukemia virus integration regions within 100 kilobases upstream of c-myb are proximal to the 5' regulatory region of the gene through DNA looping." Journal of virology **86**(19): 10524-10532.

Chapter II

A tunable zinc finger-based framework for Boolean logic computation in mammalian cells

Jason J. Lohmueller[1,2], Thomas Z. Armel[1] & Pamela A. Silver[1,2]

[1] *Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA*

[2] *Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts 02115, USA*

Reproduced from Lohmueller JJ, Armel TZ, Silver PA. (2012). A tunable zinc finger-based framework for Boolean logic computation in mammalian cells. *Nucleic Acids Res.* Jun;40(11):5180-7.

Jason J. Lohmueller contributed all data for Figures 2.1-2.5 in collaboration with Thomas Z. Armel on Figures 2.1, 2.2, 2.3, and 2.5.

ABSTRACT

The ability to perform molecular-level computation in mammalian cells has the potential to enable a new wave of sophisticated cell-based therapies and diagnostics. To this end, we developed a Boolean logic framework utilizing artificial Cys_2-His_2 zinc finger transcription factors (ZF-TFs) as computing elements. Artificial ZFs can be designed to specifically bind to different DNA sequences and thus comprise a diverse set of components ideal for the construction of scalable networks. We generate ZF-TF activators and repressors and demonstrate a novel, general method to tune ZF-TF response by fusing ZF-TFs to leucine zipper homodimerization domains. We describe 15 transcriptional activators that display 2-463 fold induction and 15 transcriptional repressors that show 1.3-16 fold repression. Using these ZFs we compute OR, NOR, AND, and NAND logic, employing hybrid promoters and split intein-mediated protein splicing to integrate signals. The split intein strategy is able to fully reconstitute the ZF-TFs, maintaining them as a uniform set of computing elements. Together these components comprise a robust platform for building mammalian synthetic gene circuits capable of precisely modulating cellular behavior.

INTRODUCTION

Mammalian synthetic gene circuits have the potential to directly impact human health in applications such as cell- and animal-based disease models, gene therapies, and smart therapeutics. Significant progress has been made to generate such networks, including: RNAi-based systems to interrogate siRNAs, microRNAs and synthetic transcription factors, RNA aptamer systems to sense multiple small molecules and metabolites, and transcription factor-based systems to sense synthetic small molecules (Kramer, Fischer et al. 2004; Rinaudo, Bleris et al. 2007; Win and Smolke 2008; Leisner, Bleris et al. 2010; Nissim and Bar-Ziv 2010; Xie, Wroblewska et al. 2011). However, despite the efficacy of these systems several engineering challenges remain. The majority of published systems rely on the use of a small set of specialized factors such as LacI, Gal4, and TetR and are not amenable to the design of large-scale networks. Methods to tune system responses over a wide dynamic range are lacking and are often specific to specialized system components. Finally, many signal integration strategies are limited by their input modularity and are not amenable to the facile generation of different types of logic gates. In this work we address these issues by developing a computational platform consisting of Cys_2-His_2 ZF-TF computing elements, general strategies to tune these ZF-TFs, and general strategies to integrate transcriptional input signals.

Cys_2-His_2 zinc fingers are small protein domains sharing a common zinc atom coordinating structural motif, many of which are capable of binding to specific DNA sequences with high affinities. Recent advances in the ability to engineer ZF DNA binding domains (DBDs) to recognize new nucleotide sequences have made them a rich potential source of independently - functioning system components (Maeder, Thibodeau-

Beganny et al. 2008; Sander, Dahlborg et al. 2011). As most artificial ZFs are constructed from a single canonical ZF, Zif268, they are also similar in amino acid composition and structure and are likely to share similar biological properties. In our systems we use a set of five previously developed ZF DBDs that target three orthogonal 9bp DNA sequences with high specificity: BCR_ABL-1, BCR_ABL-2, erbB2, HIV-1, and HIV-2 (Hurt, Thibodeau et al. 2003).

Because biological signals vary greatly in strength and composition it is desirable to have a set of tunable computing elements to interrogate these signals. A common cellular strategy to modulate transcription factor activity is through cooperative binding between TFs, a process often mediated by leucine zipper (LZ) protein-protein interaction domains (Burz, Rivera-Pomar et al. 1998). While the effect of LZs on the structure and function of ZF DBDs has been characterized biochemically, the use of LZs to modulate ZF-TF activity in cells has not been explored (Pomerantz, Wolfe et al. 1998; Wolfe, Ramm et al. 2000; Wolfe, Grant et al. 2003). We chose two well-characterized homodimerizing LZ domains with different binding affinities, the protein interaction domains of human c-Jun (K_d=448μM) and *S. cerivisiae* GCN4 (K_d=8nM), to tune our engineered ZF-TFs (Zitzewitz, Bilsel et al. 1995; d'Avignon, Bretthorst et al. 2006). We combine and compare this new tuning strategy to a previously shown method to modulate TF activity, altering the number of binding sites in ZF response promoters, further increasing the range of system tunability.

To date, relatively few transcriptional logic gates have been demonstrated, and gates have largely been pursued outside the context of a general computational framework. To compute OR and NOR logic using transcriptional networks previous

researchers have used hybrid transcription factor response promoters containing binding sites for multiple TFs (Kramer, Fischer et al. 2004). We demonstrate that it is also possible to create and tune ZF-TF based OR and NOR gates using this architecture. Previous efforts to create individual AND gates have relied on the association of split TF fragments in 2-hybrid systems, potentially limiting the activation strength of these systems and scalability (Nissim and Bar-Ziv 2010). To our knowledge, a NAND gate has not been constructed using split transcription factors in mammalian cells.

We chose to pursue split intein-mediated protein splicing as an attractive alternative approach to perform both AND and NAND computations. Split inteins, when fused to separate protein fragments, auto-catalytically splice the two fragments into a single protein without leaving a peptide scar. They have been shown to function at high efficiency in mammalian cells and in a wide range of proteins, and over 500 inteins have been discovered to date (Liu and Hu 1997; Perler 2002; Li, Sun et al. 2008). We chose to use the dnaB mini-intein from *Rhodothermus marinus* as it has been previously demonstrated to display near 100% splicing efficiency in mammalian cells (Li, Sun et al. 2008). Utilizing this protein splicing strategy we can reconstitute ZF-TFs for AND and NAND gates allowing logic computations to be enacted by complete factors, potentially yielding stronger and more uniform system responses.

MATERIALS AND METHODS

Recombinant DNA constructs

Constructs encoding zinc finger DBDs were codon-optimized for mammalian expression and synthesized (Genscript). The *rma* intein fragments and the GCN4 leucine

zipper sequences were codon optimized for mammalian expression and synthesized (Integrated DNA Technologies). Intein-zinc finger fusion parts were cloned using PCR and BbsI Type-IIS restriction enzyme methods. All experimental DNA constructs were generated by combining BioBrick subparts using Biobrick assembly (Knight 2003; Phillips and Silver 2006). Each final construct and its constituent Biobrick subparts is listed in Appendix I. Sequences of all Biobrick subparts are listed in Appendix I. For expression constructs, coding regions cut with XbaI and NotI were cloned into the NheI and NotI sites of a modified version of pCDNA5/FRT/TO (Invitrogen), "pCDNA5insVector." pCDNA5insVector was generated by cloning the subpart "pCDNA5ins" between the two PmeI sites of pCDNA5/FRT/TO. All reporter constructs were cut with SpeI and NotI and cloned between the SpeI and NotI sites of pCDNA5/FRT/TO.

Cell culture

The human osteosarcoma-derived epithelial cell line U-2 OS (ATCC #HTB-96) was maintained at 37°C, 5% CO_2 in growth medium (McCoy's 5A medium supplemented with 10% FBS, 2 mM L-glutamine, 100 U/ml penicillin and 100 µg/ml streptomyocin). A summary of the plasmid amounts used for transfections can be found in Appendix I. All transfections were performed in 12-well plates seeded with approximately 150,000 cells using 3 µl Lipofectamine LTX transfection reagent and 1 µl PLUS reagent (Invitrogen) with 1µg total DNA per well in 1 ml of growth medium. Transfection reagent was washed out and replaced with fresh growth media 6 hours post transfection.

Microscopy

Microscopy was performed on live cells in glass-bottomed wells (MatTek) in phenol red-free growth medium 48 hours post-transfection. Cells were imaged by a Nikon TE-2000 microscope with a 20x PlanFluor NA = 0.5, DIC M/N2 objective and an ORCA-ER charge-coupled device camera. Data collection and processing were performed using Metamorph 7.0 software (Molecular Devices). All images within a given experimental set were collected using the same exposure times, averaged over 3 frames, and underwent identical processing.

Flow cytometry

Approximately 30,000 live cells from each transfected well were analyzed using an LSRII cell analyser (BD Biosciences) in three biological replicates. Cells were trypsinized with 0.1 ml of 0.25% trypsin-EDTA, pelleted, and resuspended in 100 µl of Dulbecco's phosphate buffered saline containing 0.1% FBS. Output was assayed 48 hours post-transfection. First, the total CFP signal of mCh+ cells was calculated by multiplying the frequency of CFP+ cells in the mCH+ population by the mean CFP signal of these double positive cells. Fold change was calculated by dividing the total CFP of mCH+ experimental cells by the total CFP values of mCH+ off-target control cells. Values were averaged over three replicates and standard deviations were determined. In assays containing multiple off-target experiments, fold changes were calculated using the average of all off-target control wells to compare the background leakiness of the different reporter constructs.

RESULTS

Tuning ZF transcription factors

We first generated a set of 15 ZF transcriptional activators and demonstrated the ability to increase their activity with increasing LZ strength. Each activator was comprised of a ZF DBD, either no LZ motif, a Jun LZ, or a GCN4 LZ, the synthetic transcriptional activator VP64, and the SV40 nuclear localization signal (NLS). Activators were expressed from the CMV promoter and tagged with co-translationally cleaved t2a:mCherry to monitor expression (Figure 2.1A).

We assayed activator function by co-transfection with reporter plasmids containing different numbers of ZF binding sites driving expression of AmCyan fluorescent protein (CFP) (Figure 2.1A). We first tested the BCR_ABL-1 activators and observed a wide range of signal output (2-163 fold) that increased with both the strength of the LZ binding domain and the number of ZF binding sites (Figure 2.1B, Appendix I). We then compared the activity of ZF activators generated using different DBDs by co-transfection with the corresponding 6x BS reporter (Figure 2.1C, Appendix I). A strong induction from all ZFs was observed (up to 463 fold), with HIV TFs displaying the strongest activation. No cross-reactivity was observed with activators co-transfected with off-target reporters, demonstrating the specificity of our ZF-TFs (Appendix I).

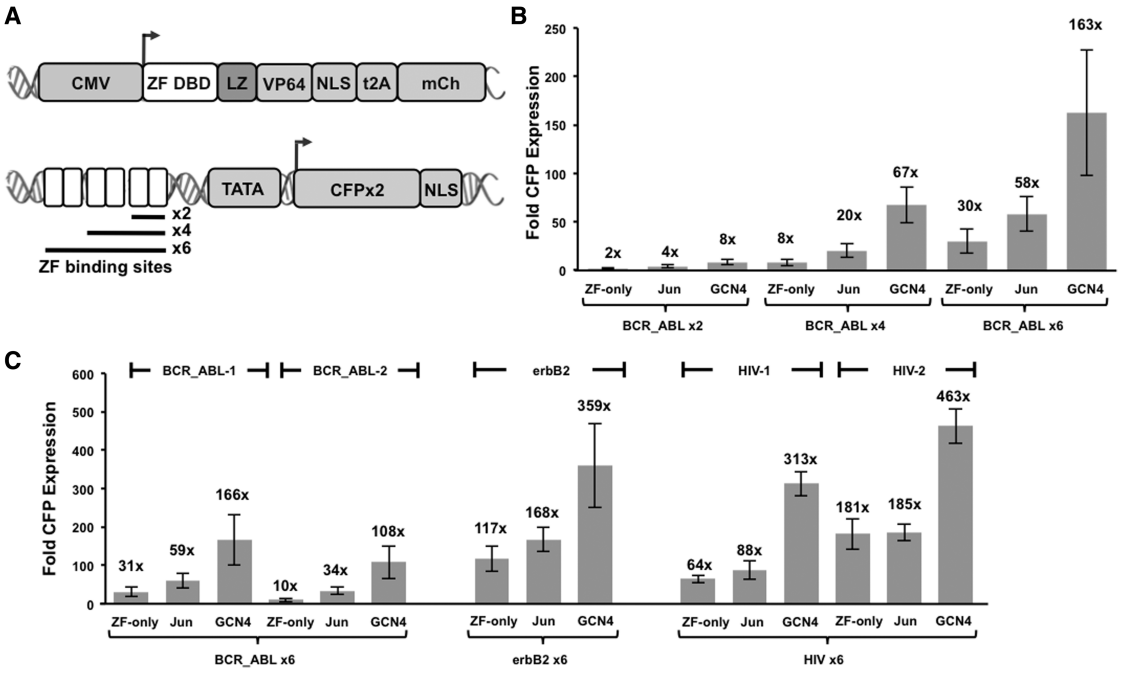


Figure 2.1 Engineering and characterization of ZF transcriptional activators

(A) Schematic representation of the assay used to test ZF activator function. Each transcriptional activator is expressed from the CMV promoter and tagged with a co-translationally cleaved mCherry fluorescent protein to monitor expression. ZF activator function is measured by the ability to activate the expression of cyan fluorescent protein from a reporter containing a minimal promoter and a variable number of 9bp ZF target sites. (B) Characterization of the role of leucine zipper addition and target site copy number on ZF-TF transcriptional activation. BCR_ABL-1 activators fused to either no LZ (ZF-only), the c-Jun LZ (Jun), or the GCN4 LZ (GCN4) were co-transfected into U-2OS cells along with reporter plasmids containing either 2, 4, or 6 copies of the corresponding 9bp target site. CFP reporter expression as measured by flow cytometry and expressed as fold change over an off-target expression control. (C) Functional characterization of all ZF-activators co-transfected with reporter plasmids containing 6 copies of their 9bp target sites.

The modular nature of our ZF elements allowed for the facile construction of a set of transcriptional repressors. These repressors can be tuned by altering the LZ dimerization domain and number of target sites analogous to our ZF activator constructs. We created 15 artificial ZF repressors by combining ZF DBDs, the SV40 NLS, and LZs with the Krüppel-associated Box (KRAB) transcriptional repression domain (Figure

2.2A). To assay for repression, we generated CFP reporter constructs containing variable

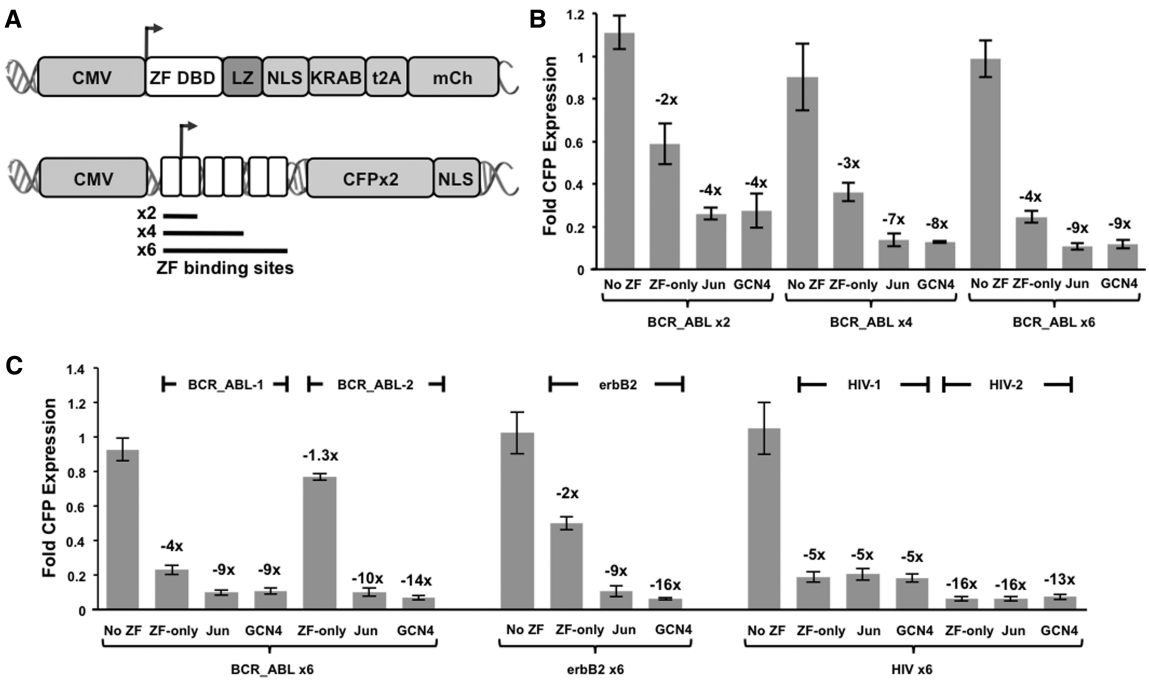


Figure 2.2 Engineering and characterization of ZF transcriptional repressors

(A) Schematic representation of the assay used to test ZF repressor function. Each transcriptional repressor is expressed from the CMV promoter and tagged with a co-translationally cleaved mCherry fluorescent protein to monitor expression. ZF repressor function is measured by the expression of cyan fluorescent protein from a CMV promoter engineered to have a variable number of 9bp ZF target sites inserted into the transcriptional start site. (B) Functional characterization of the role of target site copy number and leucine zipper addition on ZF repressor activity. BCR_ABL-1 activators fused to either no LZ (ZF-only), the c-Jun LZ (Jun), or the GCN4 LZ (GCN4) were co-transfected into U-2OS cells along with reporter plasmids containing either 2, 4, or 6 copies of the corresponding 9bp target site. The activity of each ZF repressor was determined by CFP expression measured by flow cytometry and expressed as fold change over an off-target expression control. (C) Functional characterization of all ZF-repressors co-transfected with reporter plasmids containing 6 copies of their 9bp target sites.

numbers of copies of a 9bp ZF binding site directly downstream of the TATA box within the CMV promoter (Figure 2.2A). We first tested BCR_ABL-1 repressors and observed a significant decrease in output signal (2-9 fold), strengthening with the number of ZF binding sites and the presence of a LZ (Figure 2.2B, Appendix I). All ZF repressors were

then tested with corresponding 6x reporters, and displayed a similar pattern - higher levels of repression for ZF repressors with a LZ domain (Figure 2.2C, Appendix I). The HIV repressors were the only exception displaying no LZ-mediated increase. As expected, no repression was seen for off-target reporters (Appendix I).

ZF-based Boolean logic computation

Using this set of synthetic ZF-TFs, we next sought to construct a set of Boolean logic gates. We divided the logic architecture into three general components: (i) a sensory module that senses inputs and converts the signals into ZF expression, (ii) a computational module comprised of ZFs and corresponding response promoters, and (iii) an output module consisting of binding sites controlling expression of a given protein.

Within this framework we began by generating response constructs that exhibit OR gate behavior. OR gates were developed by utilizing hybrid promoters consisting of various copies of binding sites for two distinct ZF DBDs. To determine the effect of binding site architecture on signal output, constructs containing either 2, 4, or 6 copies of the BCR_ABL and erbB2 binding sites were generated in multiple configurations (Figure 2.3A). Each configuration was tested by co-transfection with either BCR_ABL-1:GCN4 activator, erbB2:Jun activator, or both activators in tandem. All promoter architectures functioned as OR gates, with either a single activator alone or both factors present together resulting in signal output. The presence of both activators in tandem resulted in an additional 7 fold increase in output signal, likely due to the increased occupancy of reporter target sites by the transcriptional activators (Figure 2.3A). We also observed minor position effects, with activator constructs having corresponding binding sites more proximal to the TATA box displaying higher induction (Appendix I). The varying

outputs for different promoter architectures allow for an additional potential layer of tunability.

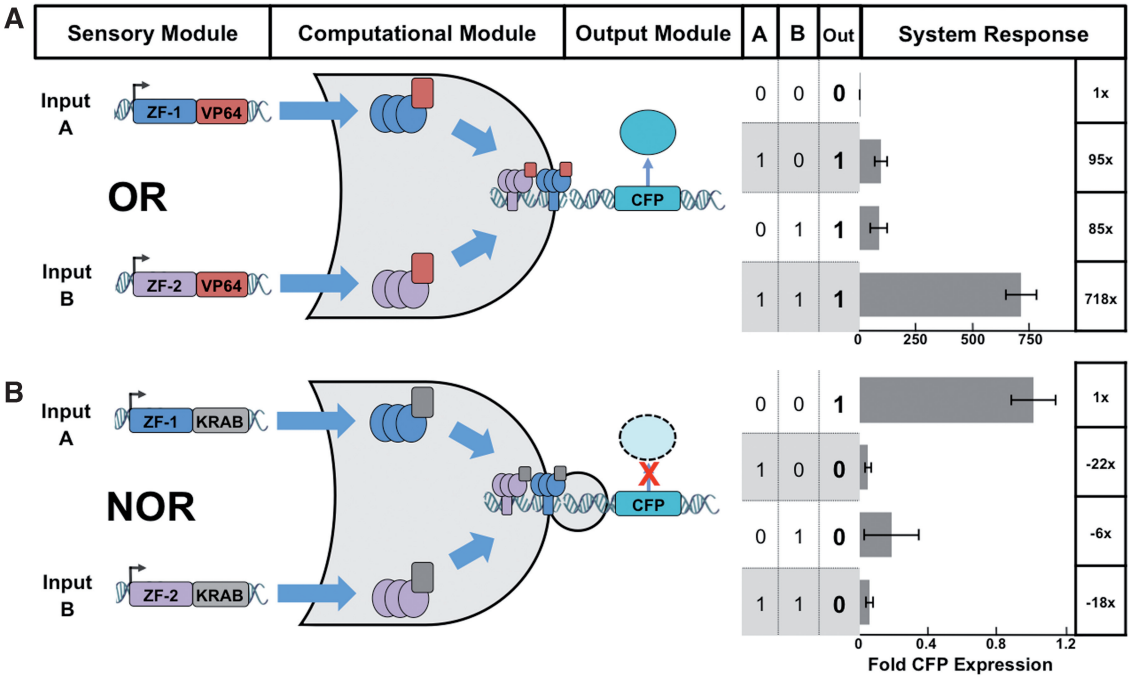


Figure 2.3 Engineering and characterization of ZF-based OR and NOR Boolean logic gates

In the sensory module, input signals lead to expression of corresponding ZF-based transcription factors. In the computational module, transcription factors act on response promoters. (A) OR gate response promoters contain target sites for two different ZF activators, and the logical operation is computed as TRUE (CFP expression) when either one or both inputs is present. For the response data shown BCR_ABL-1:GCN4 and erbB2:Jun activators were used as ZF-1 and ZF-2, respectively, and the response promoter contains 6 copies of the BCR_ABL target site upstream of 6 copies of the erbB2 target site. CFP expression was measured by flow cytometry and expressed as fold change over an off-target expression control. (B) NOR gate response promoters contain the binding sites for two different ZF repressors, and the logical operation is computed as TRUE when neither input is present. For the response data shown BCR_ABL-1:GCN4 and erbB2:Jun repressors were used as ZF-1 and ZF-2, respectively, and the response promoter contains 6 copies of the BCR_ABL target site upstream of 6 copies of the erbB2 target site. CFP expression was measured by flow cytometry.

A similar approach was employed using ZF repressors to compute NOR logic. As with our OR gates, NOR gates were generated by placing binding sites for two different ZFs within the repressor reporter plasmids. Reporters were constructed with either 2, 4,

or 6 binding sites in multiple configurations, and tested by transfecting with either BCR_ABL-1:GCN4 repressor, erbB2:Jun repressor, or both repressors. All promoter architectures functioned as NOR gates, with repression levels ranging from 2-23 fold. We again observed a minor positional effect with higher levels of repression corresponding to ZF binding site proximity to the CMV TATA box. Thus, NOR gates can also be tuned to have different response properties (Figure 2.3B, Appendix I).

We next sought to compute AND and NAND logic using our transcriptional activators and repressors. To use our split intein protein splicing strategy we first set out to determine the optimal amino acid residue at which to split our ZF-TFs. We created twelve pairs of BCR_ABL-1:Jun activator split proteins. Each pair contained an amino- (N-) and carboxy-terminal (C-) fragment fused to the appropriate intein. These fragments were co-transfected, either together or separately, with the 6x BCR_ABL activator reporter (Figure 2.4A,B,C). Fold induction was calculated relative to CFP activation by the C-terminal fragment alone. Five out of twelve zinc finger split pairs displayed greater than 3 fold signal output (Figure 2.4D, Appendix I).

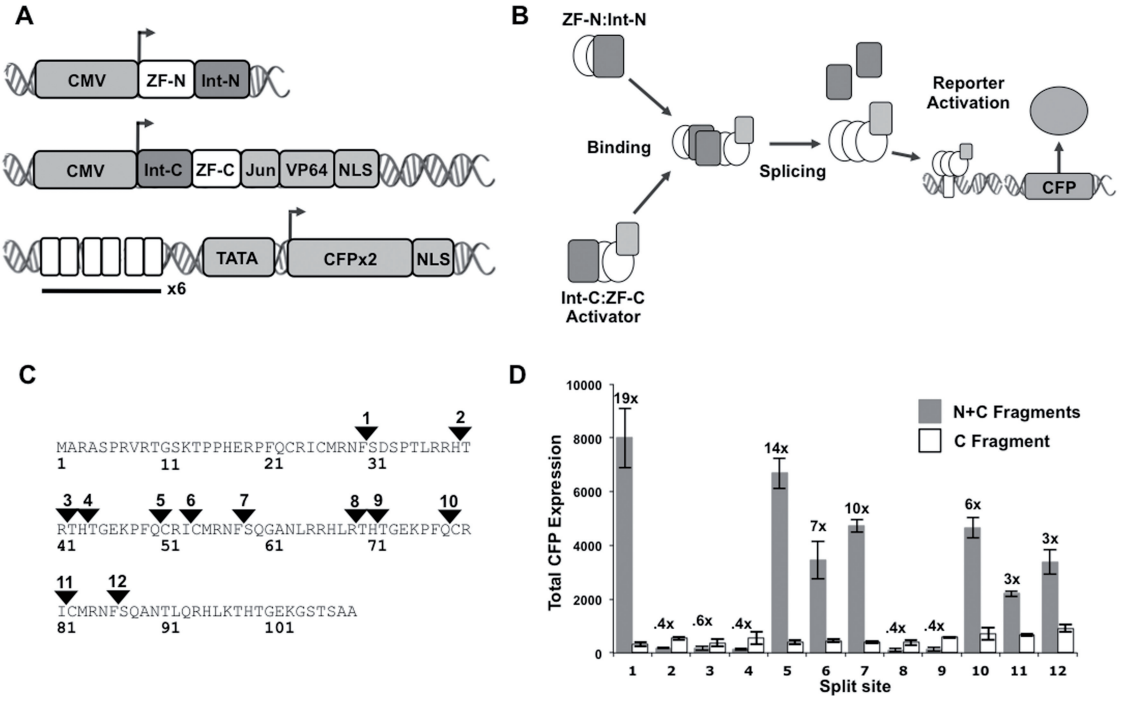


Figure 2.4 Determining the optimal ZF-TF split site

(A) Schematic representation of the plasmids used to test for ZF activator split function. Each ZF-N:int-N or Int-C:ZF-C fragment is expressed from a CMV expression plasmid. Activity is measured by the ability of each fragment pair to activate a cyan fluorescent protein reporter containing 6x copies of 9bp ZF target sites versus activation by the Int-C:ZF-C fragment alone. (B) Schematic of the ZF-TF reconstitution process. After expression, the two split ZF-intein fragments bind together and undergo protein splicing to cleave away intein fragments and reconstitute the full ZF activator leading to activation of the BCR_ABL-1 reporter. (C) BCR_ABL-1 amino acid sequence labeled with the 12 split sites assayed. (D) Characterization of the 12 ZF-intein activator pairs assayed by transient transfection in U2 OS cells. CFP reporter expression was measured by flow cytometry and reported as total CFP expression. Fold changes in total CFP are listed for the activation of each split pair compared to activation by the corresponding Int-C:ZF-C fragment alone.

The split site between residue 30 and 31 resulted in the highest activity, approximately 19-fold, and was therefore chosen for further logic gate generation. Interestingly, all functional split sites are predicted to be located in protein loop regions of the ZF, suggesting the importance of the secondary structure on efficient protein splicing (Appendix I). After performing this assay we discovered that a truncated form of

the N-terminal intein fragment, intN*, displayed higher splicing efficiency and used this intein fragment in all further experiments.

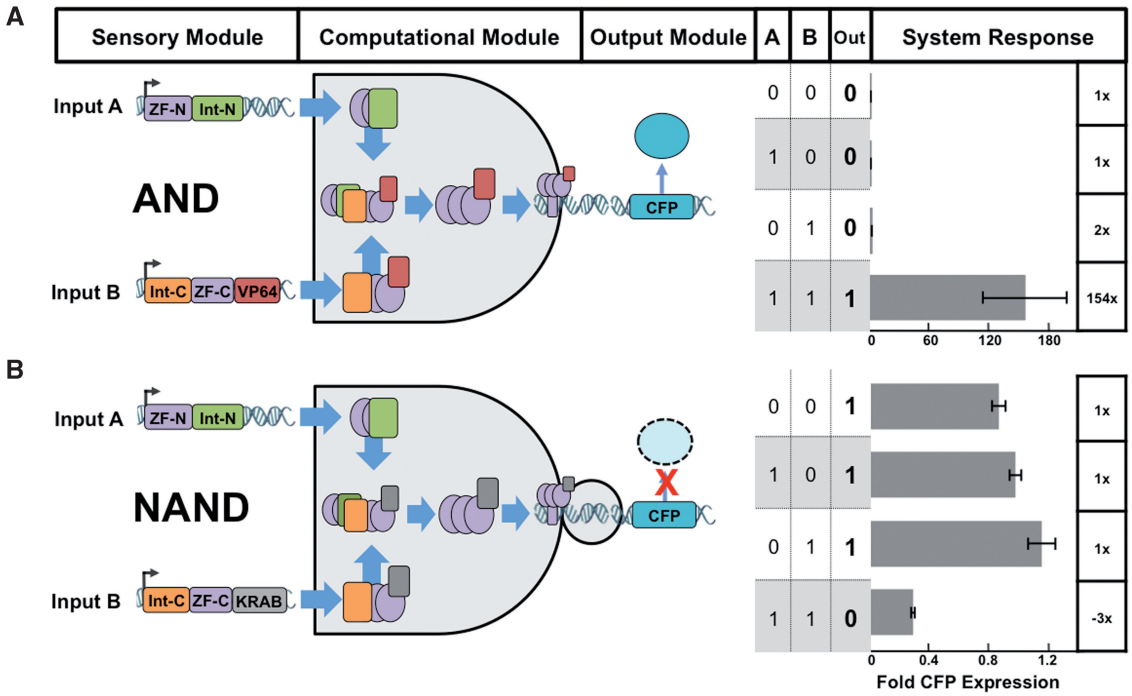


Figure 2.5. Engineering and characterization of ZF-based AND and NAND Boolean logic gates

In the sensory module, the presence of an input signal leads to the expression of the corresponding ZF-TF:split intein fragment, either ZF-N:Int-N or Int-C:ZF-C. In the computation module splicing of ZF:intein fragments leads to production of a complete ZF-TF that acts upon its cognate promoter. (A) For AND gates, a ZF activator is spliced, and the logical operation is computed as TRUE only when both input signals are present. For the response data shown BCR_ABL-1:GCN4 activator split fragments were used and the response promoter contains 6 copies of the BCR_ABL target site. CFP expression was measured by flow cytometry and expressed as fold change over an off-target expression control. (B) For NAND gates, the computational module splices a ZF repressor, and the logical operation is computed as TRUE as long as both inputs are not present together. For the response data shown BCR_ABL-1:GCN4 repressor split fragments were used and the response promoter contains 6 copies of the BCR_ABL target site. CFP expression was measured by flow cytometry and expressed as fold change over an off-target expression control.

Next, we generated an AND gate using split BCR_ABL-1:GCN4. Cells were transfected with each fragment alone or both fragments together, and an off-target ZF

DBD was used as a negative control. When both N-terminal and C-terminal fragments of BCR_ABL-1:GCN4 activator were present we observed a 154-fold signal induction, a level that is comparable to that of the parental BCR_ABL-1:GCN4 activator (Figure 2.5A, Appendix I). To compute NAND logic we sought to splice together fragments of the BCR_ABL-1:GCN4 repressor. The repressor was split in the same location as the BCR_ABL activators. We assayed for NAND activity by co-transfection of the 6x BCR_ABL repressor reporter with either fragment alone or both fragments together. No repression was seen when either the N-terminal or C-terminal fragment was transfected alone, however, when both fragments were present a 3 fold repression was observed (Figure 2.5B, Appendix I).

Next, we generated an AND gate using split BCR_ABL-1:GCN4. Cells were transfected with each fragment alone or both fragments together, and an off-target ZF DBD was used as a negative control. When both N-terminal and C-terminal fragments of BCR_ABL-1:GCN4 activator were present we observed a 154-fold signal induction, a level that is comparable to that of the parental BCR_ABL-1:GCN4 activator (Figure 2.5A, Appendix I). To compute NAND logic we sought to splice together fragments of the BCR_ABL-1:GCN4 repressor. The repressor was split in the same location as the BCR_ABL activators. We assayed for NAND activity by co-transfection of the 6x BCR_ABL repressor reporter with either fragment alone or both fragments together. No repression was seen when either the N-terminal or C-terminal fragment was transfected alone, however, when both fragments were present a 3 fold repression was observed (Figure 2.5B, Appendix I).

DISCUSSION

We developed large set of tunable ZF-TF computing elements and general transcriptional framework to perform Boolean logic operations in mammalian cells. ZF-TFs present a scalable alternative to common synthetic biology transcriptional regulators such as LacI, Gal4, and TetR with over a hundred ZF DBDs that target mutually orthogonal DNA sequences available.(7,8) We created activators and repressors using 5 previously developed ZF DBDs that target 3 orthogonal 9bp DNA sequences and corresponding response promoters. As these ZF DBDs are highly similar in structure to other artificial ZF DBDs it is likely that these too can be readily integrated into the reported system architectures. In order to generate system components that interrogate signals of different strengths and yield responses of different strengths we developed methods to tune ZF-TFs. We employed two general strategies borrowed from naturally - occurring systems: fusing ZF-TFs to LZ homo-dimerization domains and altering the number of ZF binding sites in response promoters. We created a large set of parts and elucidated general design rules based on the behaviors of these parts. Activators displayed an increase in activity correlating with both the number of promoter binding sites and the strength of the LZ interaction domain. The repressors behaved similarly displaying an increase in repression corresponding to the number of binding sites and the presence of a LZ. While the strength of the LZ domain did not impact the repression levels, this could be the result of weaker repressors reaching a maximum repression levels for the transient transfection repressor assay. Both tuning methods could be employed together to obtain maximum ZF-TF activity levels. The identity of the ZF DBD also affected ZF-TF activity, potentially due to differences in DNA binding strength or

subtle differences in ZF-TF expression levels. This element provides yet another way to tune ZF-TF. While the precise mechanism of LZ-enhanced ZF-TF activity is unclear, we speculate that it could be the result of cooperative transcription factor binding or LZ-mediated recruitment of additional ZF-TFs to the promoter region. Both mechanisms would lead to increased promoter occupancy by the ZF-TFs and a corresponding increase in activity.

To integrate signals for logic computations we employed hybrid promoter and novel split intein protein splicing integration architectures. We observed strong ON/OFF ratios for idealized CMV-expressed inputs for OR, NOR, AND, and NAND gates. The hybrid promoter strategy used for OR and NOR computations effectively integrated ZF-TF signals and showed tunability dependent on ZF binding site number and LZ strength. While all inductions were high, the OR gate behavior was somewhat analog as the presence of two inputs showed a 7 fold increase over the presence of each single input. Interestingly, this result suggests that the individual ZF-TF response promoters could perhaps be further enhanced by the addition of more ZF binding sites. The position of the ZF binding sites also affected OR gate activity with higher activation levels observed from ZF-TFs with binding sites closer to the TATA Box. The NOR gates showed comparable repression levels for single and double inputs but also displayed a position effect with stronger repression levels observed for ZF binding sites located directly on the putative transcription start sites.

To compute AND and NAND logic we combined split intein protein splicing with our ZF-TF components. Protein splicing provides a novel method to efficiently combine signals that has the advantage of generating fully formed transcription factors. We first

found the optimal split site for the ZF-TF activators among 12 putative split sites assayed. Interestingly 7 out of 12 split sites were successful at activating reporter function greater than 3 fold demonstrating the modularity of the split inteins. All successful split sites were located in protein loop regions suggesting the importance of secondary structure on splicing efficiency. The most efficient split site showed activity levels matching that of the complete factor. As the split sites are located within the constant region of the BCR_ABL1 ZF DBD it is likely they can be used in ZF-TFs containing different ZF-DBDs. Finally, the optimal ZF split site was effective in the context of ZF activators with different LZ domains and also in the context of a repressor to perform NAND logic. The NAND gate showed around a 3 fold repression of the reporter. While significantly lower than the complete factor (9-fold repression) this is possibly due to a delay caused by intein splicing that lead to leakiness from the CMV promoter before it can be efficiently repressed.

Together these system components greatly expand the repertoire of parts and devices available to mammalian synthetic biologists. The tuning methods were effective at generating a large variation in ZF-TF activities and could potentially be generalized to tune other transcription factors. While the reported systems can stand on their own the parts and architectures could also be used in conjunction with other existing logic computational methods that rely on the use of synthetic transcription factors. The logic framework developed utilizing these factors provides a powerful new and general method for computing Boolean logic in mammalian cells. Through the use of modular ZF DBDs and LZ dimerization domains we have developed an approach that should be readily scalable to different input/output requirements and tolerances. Components in the system,

while orthogonal in specificity, share common structural and functional qualities, promising to make optimization of networks more streamlined and different networks more comparable. The potential modularity of cellular inputs and the tunability of output response within this framework lends itself to the processing of multiple cellular signals and the future rewiring of intrinsic network topologies to engineer precise biological responses. Future scaling up of these systems can follow our elucidated design principles and take advantage of the many additional LZ's, ZF DBDs and split inteins available.

ACKNOWLEDGMENTS

The authors wish to acknowledge M. Hoerner for technical assistance, as well as D. Burrill, M. Inniss, A. Garg and all members of the Silver lab for helpful comments and discussion. The authors would also like to acknowledge Y.P. Hung, D.B. Thompson, C.J. Delebecque, and D. Burrill for carefully reading the manuscript. This work was supported by funds from NIH 1F32 CA154195-01 to T.Z.A., the Wyss Institute for Biologically Inspired Engineering to J.J.L. and P.A.S., and NIH R37 GM36373-22 to P.A.S.

REFERENCES

Burz, D. S., R. Rivera-Pomar, et al. (1998). "Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent gene activation in the Drosophila embryo." The EMBO journal **17**(20): 5998-6009.

d'Avignon, D. A., G. L. Bretthorst, et al. (2006). "Site-specific experiments on folding/unfolding of Jun coiled coils: thermodynamic and kinetic parameters from spin inversion transfer nuclear magnetic resonance at leucine-18." Biopolymers **83**(3): 255-267.

Hurt, J. A., S. A. Thibodeau, et al. (2003). "Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection." Proceedings of the National Academy of Sciences of the United States of America **100**(21): 12271-12276.

Knight, T. (2003). "Idempotent Vector Design for Standard Assembly of Biobricks." DSpace. MIT Artificial Intelligence Laboratory;MIT Synthetic Biology Working Group.

Kramer, B. P., C. Fischer, et al. (2004). "BioLogic gates enable logical transcription control in mammalian cells." Biotechnology and bioengineering **87**(4): 478-484.

Leisner, M., L. Bleris, et al. (2010). "Rationally designed logic integration of regulatory signals in mammalian cells." Nature nanotechnology **5**(9): 666-670.

Li, J., W. Sun, et al. (2008). "Protein trans-splicing as a means for viral vector-mediated in vivo gene therapy." Human gene therapy **19**(9): 958-964.

Liu, X. Q. and Z. Hu (1997). "A DnaB intein in Rhodothermus marinus: indication of recent intein homing across remotely related organisms." Proceedings of the National Academy of Sciences of the United States of America **94**(15): 7851-7856.

Maeder, M. L., S. Thibodeau-Beganny, et al. (2008). "Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification." Molecular cell **31**(2): 294-301.

Nissim, L. and R. H. Bar-Ziv (2010). "A tunable dual-promoter integrator for targeting of cancer cells." Molecular systems biology **6**: 444.

Perler, F. B. (2002). "InBase: the Intein Database." Nucleic acids research **30**(1): 383-384.

Phillips, I. and P. A. Silver (2006) "A New Biobrick Assembly Strategy Designed for Facile Protein Engineering." DSpace. MIT Artificial Intelligence Laboratory;MIT Synthetic Biology Working Group.

Pomerantz, J. L., S. A. Wolfe, et al. (1998). "Structure-based design of a dimeric zinc finger protein." Biochemistry **37**(4): 965-970.
Rinaudo, K., L. Bleris, et al. (2007). "A universal RNAi-based logic evaluator that operates in mammalian cells." Nature biotechnology **25**(7): 795-801.

Sander, J. D., E. J. Dahlborg, et al. (2011). "Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA)." Nature methods **8**(1): 67-69.

Win, M. N. and C. D. Smolke (2008). "Higher-order cellular information processing with synthetic RNA devices." Science **322**(5900): 456-460.

Wolfe, S. A., R. A. Grant, et al. (2003). "Structure of a designed dimeric zinc finger protein bound to DNA." Biochemistry **42**(46): 13401-13409.

Wolfe, S. A., E. I. Ramm, et al. (2000). "Combining structure-based design with phage display to create new Cys(2)His(2) zinc finger dimers." Structure **8**(7): 739-750.

Xie, Z., L. Wroblewska, et al. (2011). "Multi-input RNAi-based logic circuit for identification of specific cancer cells." Science **333**(6047): 1307-1311.

Zitzewitz, J. A., O. Bilsel, et al. (1995). "Probing the folding mechanism of a leucine zipper peptide by stopped-flow circular dichroism spectroscopy." Biochemistry **34**(39): 12812-12819.

Chapter III

Engineering synthetic TAL effectors with orthogonal target sites

Abhishek Garg[1], Jason J Lohmueller[1], Pamela A Silver[1,2], and Thomas Z Armel[1]

[1]*Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA*

[2]*Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts 02115, USA*

Reproduced from Garg A, Lohmueller JJ, Silver PA, Armel TZ. (2012). Engineering synthetic TAL effectors with orthogonal target sites..*Nucleic Acids Res.* Aug;40(15):7584-95.

Jason J. Lohmueller contributed data for Figures 3.1,3.5-8 in collaboration with Abhishek Garg, and contributions from Thomas Z. Armel on Figures 3.5 and 3.6. Abhishek Garg contributed Figures 3.2 and 3.3.

ABSTRACT

The ability to engineer biological circuits that process and respond to complex cellular signals has the potential to impact many areas of biology and medicine. Transcriptional activator-like effectors (TALEs) have emerged as an attractive component for engineering these circuits, as TALEs can be designed *de novo* to target a given DNA sequence. Currently, however, the use of TALEs is limited by degeneracy in the site-specific manner by which they recognize DNA. Here, we propose an algorithm to computationally address this problem. We apply our algorithm to design 180 TALEs targeting 20 bp cognate binding sites that are at least 3 nucleotide mismatches away from all 20 bp sequences in putative 2kb human promoter regions. We generated 8 of these synthetic TALE activators and showed that each is able to activate transcription from a targeted reporter. Importantly, we show that these proteins do not activate synthetic reporters containing mismatches similar to those present in the genome nor a set of endogenous genes predicted to be the most likely targets *in vivo*. Finally, we generated and characterized TALE repressors comprised of our orthogonal DNA binding domains and further combined them with shRNAs to accomplish near complete repression of target gene expression.

INTRODUCTION

A central goal of synthetic biology is the creation of gene regulatory circuits that specifically and robustly control gene expression in response to cell state and environmental cues (Andrianantoandro, Basu et al. 2006; Haynes and Silver 2009; Tabor, Salis et al. 2009; Khalil and Collins 2010). While much progress has been made toward developing genetic systems that detect biological signals, the ability to integrate these signals has been limited by the lack of modular and mutually orthogonal genetic elements available for use. Additionally, the functionality of these systems can be hampered by unwanted interference with the host cell machinery. The generation of high-fidelity gene circuits would thus benefit from a set of mutually orthogonal synthetic regulatory components that have minimal effects on endogenous cell machinery. In the case of transcriptional systems it would be ideal to have a set of transcriptional regulators that would only target DNA sequences that exist within the artificial circuit. Such regulators would have minimal affinity for DNA sequences present in the endogenous promoter regions of the host cell, thus minimizing unwanted effects on host gene expression (Figure 3.1). Transcription factors with programmable DNA binding domains offer one potential approach toward this goal. Transcription activator-like effector (TALE) proteins have been recently demonstrated to have modular and predictable DNA binding domains, thereby allowing for the *de novo* creation of synthetic transcription factors that bind any DNA sequence of interest (Christian, Cermak et al. 2010; Morbitzer, Romer et al. 2010; Li, Huang et al. 2011; Miller, Tan et al. 2011; Deng, Yan et al. 2012; Mak, Bradley et al. 2012).

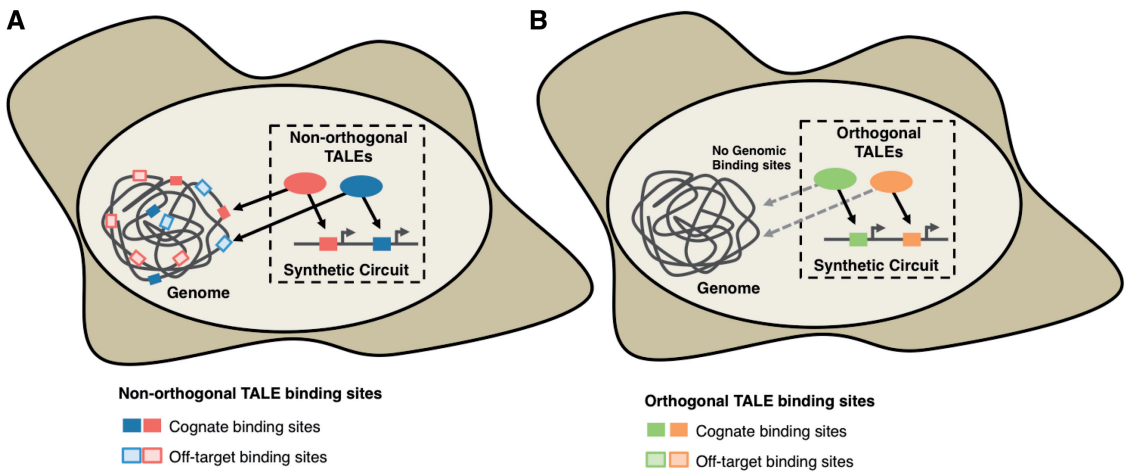
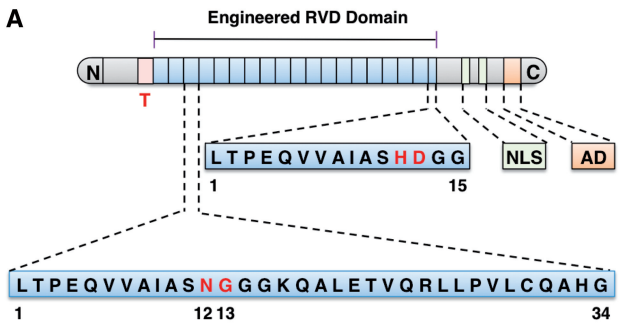


Figure 3.1 Orthogonal TALEs as ideal regulatory components for insulated synthetic gene circuits

(A) Non-orthogonal TALEs designed to bind and regulate gene expression of a synthetic gene circuit may also bind to cognate and off-target (containing mismatches) binding sequences in the endogenous promoter regions in the genomic DNA. (B) Orthogonal TALEs bind and regulate gene expression of a synthetic gene circuit and have no predicted binding sites in the endogenous promoter regions.

Originally discovered in phytopathogenic bacteria of the genus *Xanthomonas,* TALE proteins are made up of three distinct regions; (i) an N-terminal region housing the protein secretion and translocation signals, (ii) a central repeat domain composed of a series of tandem repeats containing repeat variable di-residues (RVDs) that specifically recognize and bind DNA, and (iii) a C-terminal domain containing two nuclear localization signals (NLSs) and a transcriptional activation domain (Figure 3.2A) (Herbers, Conrads-Strauch et al. 1992; Van den Ackerveken, Marois et al. 1996; Zhu, Yang et al. 1998; Boch and Bonas 2010). The central DNA binding domain is composed of a variable number of 33 - 35 amino acid repeats such that each binding domain recognizes a different DNA base pair (bp) and can be recombined to recognize any given DNA sequence (Boch and Bonas 2010; Scholze and Boch 2011). Recent studies have deciphered the code by which the repeat elements bind to DNA, showing that the

residues at amino acid positions 12 and 13 in each repeat determine which nucleotide is preferentially bound (Figure 3.2B) (Boch, Scholze et al. 2009; Moscou and Bogdanove 2009). The modularity of these repeat elements has enabled TALEs to become a powerful tool, allowing for the creation of synthetic transcriptional activators that can target a specific DNA sequence and activate a desired gene (Morbitzer, Romer et al. 2010; Miller, Tan et al. 2011). Furthermore, because of the protein's modular nature, TALEs are amenable to hierarchical ligation-based construction strategies, enabling the development of large libraries of proteins (Cermak, Doyle et al. 2011; Li, Huang et al. 2011; Miller, Tan et al. 2011; Morbitzer, Elsaesser et al. 2011; Weber, Gruetzner et al. 2011).



B

RVD	Amino acid sequence	Preferred Target Nucleotide
NI	LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG	A
HD	LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG	C
NG	LTPQQVVAIASNGGGKQALETVQRLLPVLCQAHG	T
NK	LTPDQVVAIANNKGGKQALETLQRLLPVLCQAHG	G

Figure 3.2 TALE protein architecture and DNA binding specificities

(A) Schematic of a representative TALE protein with 18.5 Repeat Variable Di-residue (RVD) domains. Each RVD domain is composed of 34 amino acids and differs only in the variable amino acids highlighted in red. The C-terminal RVD domain is a 15 amino acid half repeat domain. The 2 endogenous NLS domains and the endogenous activation domain (AD) present in naturally occurring TALEs were replaced with SV40 NLSs and the VP64 activation domain, respectively. (B) The amino acid sequences of the AvrBs3 NI, HD, NG, and NK RVD domains and their preferred target nucleotides.

At present, however, drawbacks to the use of TALEs as targeted transcription factors exist. Most notably, each TALE repeat does not bind to a given DNA base pair with perfect complementarity (Figure 3.2B) (Romer, Recht et al. 2009; Scholze and Boch 2010). While it has been shown that in some cases including a single mismatch in the binding site of a given TALE can significantly inhibit its off-target activity, there are known instances where designed TALEs have been demonstrated to bind to unintended off-target DNA sequences that differ from their cognate target sequence by up to 3bp (as defined by the TALE binding code) (Boch, Scholze et al. 2009; Moscou and Bogdanove 2009). These observations indicate that while a synthetic TALE can be designed to efficiently target a given DNA sequence, unintended off-target effects can frequently occur and may limit the utility of TALEs for specifically controlling the expression of a targeted gene (Appendix II). This limitation also restricts the application of TALEs as components of synthetic circuits where orthogonality to the host cell's genome is an important constraint.

We have developed an algorithm that allows one to computationally design TALEs with cognate binding sites that are at least a given number of mismatches away from a set of DNA sequences. We apply our algorithm to design TALEs with 20 bp cognate binding sites that are at least 3 nucleotide mismatches away from all 2000 bp putative human promoter sequences and at least 4 nucleotide mismatches from 500 bp putative human promoter sequences. These TALEs represent a potentially powerful set of insulated transcriptional regulators for the construction of synthetic gene circuits. We generated DNA constructs encoding 8 of these TALEs as transcriptional activators and assessed their function in human cells. We demonstrate that each TALE effectively

activates transcription from its targeted binding site and that the TALE activators are mutually orthogonal in their activities. We also show that the TALEs do not activate transcription from artificial promoters containing binding sites comparable to potential off-target sites in human promoter regions and provide additional evidence that the TALEs do not activate their closest off-target endogenous genes. Finally, we use two of the TALE DNA binding domains to generate TALE repressors and demonstrate strong TALE-mediated repression of a reporter gene. We further combine these TALE repressors with synthetic shRNAs targeting the same reporter to obtain even stronger, near complete gene repression. Our methodologies and TALE transcription factors address a major gap in synthetic biology and provide a new set of tools toward the design of robust genetic circuits that function orthogonally to the cells in which they are utilized (An and Chin 2009; Lu, Khalil et al. 2009; Barrett and Chin 2010; Wang, Kitney et al. 2011).

MATERIAL AND METHODS

Human Genome DNA sequences

The sequences corresponding to the 2000 bp regions upstream of all annotated transcription start sites in human RefSeq genes with annotated 5' untranslated regions (UTRs) were downloaded from the UCSC Genome browser website (http://genome.ucsc.edu/). If multiple upstream regions per RefSeq gene were found due to multiple annotates transcription start sites, then all upstream regions were used for computing orthogonal 20-mers. Downloaded sequence files correspond to the Feb. 2009

assembly of the human genome (hg19, GRCh37 Genome Reference Consortium Human Reference 37).

Recombinant DNA constructs of TALEs and reporters

Amino acid sequences encoding all TALE constructs were derived from the AvrBs3 amino acid sequence (GenBank locus id. CAA34257.1), including sequences encoding the sub-modules corresponding to the constant 5' region, variable repeats regions (for di-residues HD, NI and NG) and the constant 3' region. Within these sequences the naturally existing NLS regions and activation domains in AvrBs3 were identified in the 3' constant region and replaced with mammalian SV40 NLS and VP64 activation domains. For TALE repressors the VP64 activation domain was replaced with the KRAB transcriptional repression domain. DNA sequences encoding these components were codon-optimized for expression in human cells and synthesized by Integrated DNA Technology (Coralville, IA). The exact positions and sequences used are listed in Appendix II. These components were combined to generate TALE expression constructs using a hierarchal cloning scheme outlined in Appendix II. t2A and mCherry were combined to full length TALE activator coding regions and t2A and DsRed-shRNA consturcts were combined to full length TALE repressor coding regions using BioBrick cloning. These complete coding regions were cloned into the NheI and NotI sites of pCDNA5insVector for expression from the CMV promoter (Knight 2003; Phillips 2006; Lohmueller, Armel et al. 2012).

Reporter constructs for activators and repressors were cloned using BioBrick assembly, cut with SpeI and NotI, and cloned between the SpeI and NotI sites of pCDNA5/FRT/TO for mammalian expression (Invitrogen, Carlsbad, CA). Finally, to

create combined TALE repressor and shRNA reporter constructs, shRNA target sites FF4' and FF6' were cloned into the NotI site of the CFP reporter constructs of the TALE repressors.

Cell culture

The human osteosarcoma-derived epithelial cell line U-2OS (American Type Culture Collection, Manassas, VA) was maintained at 37° C, 5% CO2 in growth medium (McCoy's 5A medium supplemented with 10% FBS, 2 mM L-glutamine, 100 U/ml penicillin and 100 ug/ml streptomyocin). The human embryonic kidney cell line HEK293 (American Type Culture Collection, Manassas, VA) was maintained at 37° C, 5% CO2 in growth medium (Dulbecco's Modified Eagle Medium supplemented with 10% FBS, 2 mM L-glutamine, 100 U/ml penicillin and 100 ug/ml streptomyocin). All transfections were performed in 12-well plates seeded with approximately 175,000 cells using 3 µl Lipofectamine LTX transfection reagent and 1 µl PLUS reagent (Invitrogen, Carlsbad, CA). All TALE activator transfections were performed in U-2OS cells and used 25 ng of TALE expression plasmid with 975 ng of reporter plasmid in 1 ml of growth medium. TALE repressor experiments were performed in HEK293 cells and used 100 ng of TALE expression plasmid with 10 ng of reporter plasmid and 890 ng of empty pCDNA5insVector in 1 ml of growth medium.

Microscopy

All microscopy was performed on live cells in glass-bottomed wells (MatTek, Ashland, MA) in phenol red-free growth medium 24 h post-transfection. Cells were imaged using a Nikon TE-2000 microscope with a 20x PlanFluor NA = 0.5, DIC M/N2

objective and collected with an ORCA-ER charge-coupled device camera. Data collection and processing were performed with Metamorph 7.0 software (Molecular Devices, Sunnyvale, CA). All images for a given experimental set and the corresponding controls were collected with the same exposure times, averaged over 3 frames, and underwent identical processing.

Flow Cytometry

Approximately 30,000 cells from each transfected well were analyzed using an LSRII cell analyzer (BD Biosciences, San Jose, CA) in three biological replicates. Cells were trypsinized with 0.1 ml of 0.25% trypsin-EDTA, pelleted, and resuspended in 100 µl of Dulbecco's phosphate buffered saline containing 0.1% FBS. For activator experiments output was assayed 24 h post-transfection. The total AmCyan fluorescent protein (CFP) signal of mCh+ cells was calculated by multiplying the frequency of CFP+ cells in the mCH+ population by the mean CFP signal of these double positive cells. The fold change of AmCyan reporter fluorescence was then calculated as the ratio of total AmCyan fluorescence intensity of cells transfected with on-target TALE expression plasmids to cells transfected with reporter plus an off-target input. For repressor experiments output was assayed 48 h post-transfection and fold change of AmCyan reporter fluorescence was calculated as the ratio of total AmCyan fluorescence intensity of DsRED+ cells transfected with on-target TALE-shRNA expression plasmids to DsRED+ cells transfected with a reporter plus an off-target input. To isolate the effects of TALEs and shRNAs, expression constructs with different combinations of TALE5R, TALE8R, FF4 and FF6 represented different on- and off-target combinations depending on the co-transfected reporter (Appendix II).

Quantitative PCR

For mRNA quantification, mCherry positive U-2OS cells were sorted and collected 48 h post transfection. RNA was extracted from mCherry positive cells using the RNeasy mini kit (Qiagen, Valencia, CA), and mRNA levels were quantified using the SYBR Green Assay (Applied Biosystems, Foster City, CA). The mRNA to cDNA conversion was performed using the SuperScript III RT kit (Invitrogen, Carlsbad, CA). Three biological replicates per sample and three technical replicates per assay were analyzed for absolute quantification of mRNA levels in transfected cells. Two biological replicates were analyzed for the mRNA levels quantification of OSGIN2 and ZC3H10 in cells transfected with TALE5 and TALE8 respectively. Relative transcript levels were assessed using the $2^{-\Delta\Delta Ct}$ method (16) with GAPDH as a reference gene. Statistical comparison between groups was made by the pairwise fixed reallocation randomization test using the publicly available Relative Expression Software Tool (REST) (Pfaffl, Horgan et al. 2002).The off-target and on-target DNA sequences of TALEs are detailed in Appendix II. Primer sequences used for qPCR are detailed in Appendix II.

Algorithm Implementation

The algorithm was implemented in C++ and the software binaries are made available for download at http://silver.med.harvard.edu/tale.html. Further details about the algorithm are provided in Appendix II. All the results presented here were obtained by running our software on the Harvard Medical School shared research cluster of computation nodes.

RESULTS

Design and implementation of an algorithm for finding orthogonal TALE binding sites

We first sought to computationally design a set of TALEs that bind to 20 bp nucleotide sequences (20-mers) and are orthogonal to human promoter regions. A TALE is defined to be orthogonal to a set of sequences if it is not predicted to bind to any sequence in that set. In this context, the number of base pair mismatches between a TALE's target sequence and a potential off-target sequence (also referred to as the *hamming distance* between the two sequences) is the main determinant of the orthogonality of the TALE. Thus, a large hamming distance between the TALE target site and a potential off-target sequence corresponds to a lower chance of the TALE binding to that off-target sequence.

To design synthetic TALEs that function orthogonally to a set of non-intended target sites, we developed an algorithm based on the *farthest string problem*. Given a set, S, of n-mers defined over an alphabet, Γ, (e.g. $\Gamma = \{A,C,G,T\}$), the objective of the farthest string problem is to find an n-mer (over the alphabet Γ) that has the largest minimum hamming distance to n-mers in set S. The farthest string problem belongs to a class of NP-hard problems for which no polynomial time solution is known to exist (Lanctot 2003). Therefore, it may take an exponential amount of time to enumerate all possible 4^{20} nucleotide sequences and test each to find a 20-mer at a maximum hamming distance from the set of genomic 20-mers. At present no algorithm exists to efficiently compute a set of such n-mers. However, by designing careful heuristics, our algorithm

can efficiently find a list of 20-mers that are orthogonal to human genome promoter regions by a hamming distance of 3 bp or more.

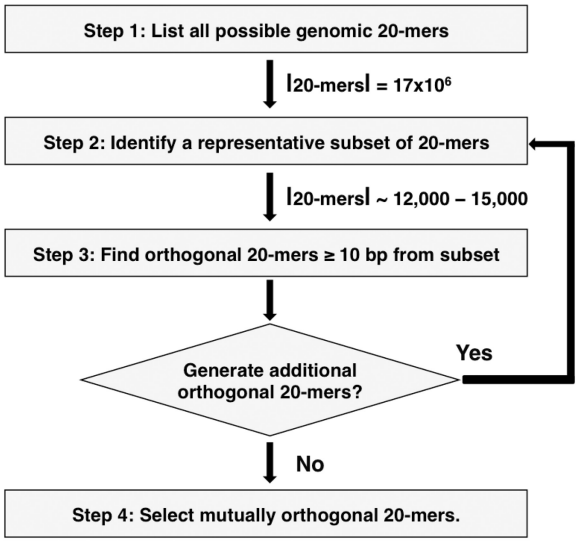


Figure 3.3 Flowchart enumerating the steps used in our algorithm to compute orthogonal 20-mers

Steps 1-2 describe the process used to reduce the set of genomic 20-mers. Steps 2 and 3 describe the process of obtaining 20-mers orthogonal to the genomic set. Steps 2 and 3 of the algorithm can be iterated until the desired number of orthogonal sequences has been computed. Finally, the resulting sets of TALEs are checked for mutual orthogonality to avoid cross-interference within the synthetic circuits.

The steps followed by our algorithm are outlined in Figure 3.3. We began by using a sliding window approach to enumerate all possible 20-mers present across both DNA strands in the promoter region of all genes in the human genome. We define promoter regions as the 2000 bp regions upstream of the transcription start site (TSS) of each gene. Because the presence of a 5' T has been demonstrated to be a necessary condition for efficient TALE binding, 20-mers that do not begin with T were not considered, yielding a total of $17x10^6$ 20-mers that are potential TALE binding sites (Boch, Scholze et al. 2009). To further reduce the number of 20-mers, the parental set of $17x10^6$ 20-mers was divided into subsets, such that each subset could be represented by a

single 20-mer within a 7 bp hamming distance from all sequences in that subset (Appendix II). Due to the reverse triangle inequality property of hamming distances, all 20-mers that are at a minimum 10 bp hamming distance from these representative sequences will also be at a minimum hamming distance of 3bp from the parental set of $17x10^6$ genomic 20-mers (Appendix II). Our algorithm uses symbolic modeling techniques and Boolean algebra to find all possible 20-mers at a minimum hamming distance of 10 bp from representative sequences of each subset (Appendix II). Multiple solutions to finding such subsets exist and each solution is typically comprised of 12,000-15,000 subsets, each having a representative 20-mers. By generating multiple sets of representative 20-mers and applying our algorithm iteratively, we identified over 180 potential binding sites for synthetic TALEs at a minimum hamming distance of 3 bp from any 20-mer in the promoter regions of the human genome (Appendix II).

We chose to generate and characterize 8 of these 180 TALEs predicted to be orthogonal to human promoter regions (Table 3.1). Chosen TALEs had a hamming distance of 3 bp from all 2000 bp genomic promoter regions and a hamming distance of 4 bp from 500 bp genomic promoter regions. The hamming distance to the more stringent 500 bp genomic promoter regions was used as an additional criterion as native transcription factor binding sites are known to be highly concentrated within these 500 bp regions proximal to the TSS (Xie, Lu et al. 2005; Carninci, Sandelin et al. 2006; Koudritsky and Domany 2008; MacIsaac, Lo et al. 2010). From our set of 150 synthetic TALEs, 100 proteins possessed a minimum hamming distance of 4 bp from 500 bp proximal promoter regions, while the remaining 50 proteins had a hamming distance of 3

bp. To minimize potential cross-activation between the selected TALEs, we also ensured that the 8 selected TALEs were predicted to be mutually orthogonal.

Table 3.1 Constituent RVDs and cognate binding sites of the 8 TALEs that were constructed and functionally characterized

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
TALE 1		HD	NI	NI	NG	NI	HD	NG	NG	NI	HD	NI	NI	NI	HD	NG	HD	HD	NG	NG
	T	C	A	A	T	A	C	T	T	A	C	A	A	A	C	T	C	C	T	T
TALE 2		HD	HD	NI	HD	HD	NI	NI	NI	NG	NG	HD	NI	NI	HD	NI	HD	NG	NG	NG
	T	C	C	A	C	C	A	A	A	T	T	C	A	A	C	A	C	T	T	T
TALE 3		HD	NI	NG	HD	NG	NI	HD	NI	NI	HD	NI	HD	NG	NI	HD	NG	NI	NG	NG
	T	C	A	T	C	T	A	C	A	A	C	A	C	T	A	C	T	A	T	T
TALE 4		HD	HD	HD	NI	NI	NG	NI	HD	NI	HD	NG	NI	NG	NI	NI	HD	NI	HD	NI
	T	C	C	C	A	A	T	A	C	A	C	T	A	T	A	A	C	A	C	A
TALE 5		NI	NI	HD	NG	NG	NI	HD	HD	NG	NG	HD	NG	HD	NI	NI	HD	NI	HD	NI
	T	A	A	C	T	T	A	C	C	T	T	C	T	C	A	A	C	A	C	A
TALE 6		NI	NG	HD	HD	NG	HD	NG	NG	NI	HD	NI	NI	NG	NI	NG	HD	HD	HD	NI
	T	A	T	C	C	T	C	T	T	A	C	A	A	T	A	T	C	C	C	A
TALE 7		NI	HD	NG	NG	NI	HD	HD	HD	NG	NI	NI	HD	HD	HD	NI	NI	NG	NG	NG
	T	A	C	T	T	A	C	C	C	T	A	A	C	C	C	A	A	T	T	T
TALE 8		NI	NG	NI	HD	NG	NI	NG	HD	HD	NI	NI	NG	HD	HD	NI	NI	HD	NG	NG
	T	A	T	A	C	T	A	T	C	C	A	A	T	C	C	A	A	C	T	T
TALE OSGIN2		HD	HD	NG	HD	HD	HD	HD	NI	HD	HD	NG	NG	NG	NI	NI	NG	NG	NG	NG
	T	C	C	T	C	C	C	C	A	C	C	T	T	T	A	A	T	T	T	T
TALE ZC3H10		NI	HD	HD	NI	NG	NI	NG	HD	HD	HD	NI	NG	HD	HD	NI	NI	HD	NG	HD
	T	A	C	C	A	T	A	T	C	C	C	A	T	C	C	A	A	C	T	C

In vivo characterization demonstrates activity and mutual orthogonality of synthetic TALE activators

To generate each of our 8 computationally designed TALEs for assaying *in vivo*, a library of subparts was synthesized containing both individual di-residue repeats and each pairwise combination of repeats, codon-optimized for expression in mammalian cells. Individual TALEs were created using a hierarchical, modular cloning strategy that leverages type IIS restriction enzymes to readily combine members of a library of

subparts into any desired TALE (Appendix II). The modular cloning scheme we use is similar to the techniques reported in the recent literature (Cermak, Doyle et al. 2011; Geissler, Scholze et al. 2011; Li, Huang et al. 2011; Morbitzer, Elsaesser et al. 2011; Weber, Gruetzner et al. 2011; Zhang, Cong et al. 2011). For each protein, both native NLSs were replaced with eukaryotic versions, and the native activation domain was replaced with the VP64 mammalian transcriptional activation domain. TALEs were expressed from the cytomegalovirus (CMV) promoter and tagged with an auto-catalytically cleaved t2A peptide fused to mCherry fluorescent protein as a transfection control (Figure 3.4A).

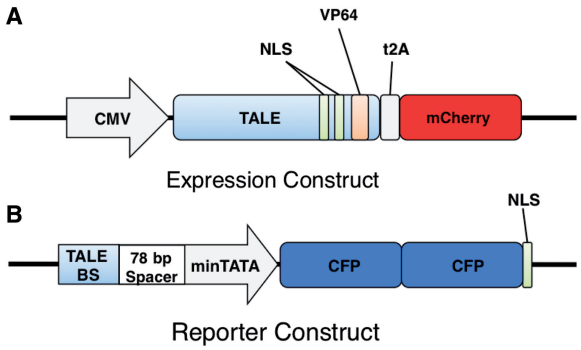


Figure 3.4 Schematic of TALE expression constructs

(A) Schematic of TALE expression constructs. Each TALE coding region was cloned into a mammalian expression vector downstream of the CMV promoter. All synthetic TALEs were also tagged with an self-cleaving t2A:mCherry fluorescent protein as a transfection control. (B) Schematic of TALE reporter constructs. Reporter constructs were generated by cloning a 20 bp TALE target sequence upstream of a minimal TATA box separated by a 78 bp spacer region. Binding of a TALE activator to the 20 bp target sequence drives expression of 2 tandem copies of NLS-tagged CFP cloned downstream of the TALE-responsive promoter as an output for TALE functionality.

The ability of our synthetic proteins to recognize a binding site and activate gene expression was tested by co-transfecting TALE expression constructs with reporter plasmids containing a 20-mer binding site driving expression of two tandem copies of the

AmCyan fluorescent protein (CFP) fused to an NLS (Figure 3.4B). Experiments were performed in the U-2OS human osteosarcoma cell line and assayed by fluorescence microscopy and flow cytometry 24 hours post-transfection. Each TALE was co-transfected with its corresponding binding site reporter plasmid to determine if it was capable of activating transcription from its targeted reporter, as well as with reporter plasmids containing binding sites for the seven other constructed TALEs in order to ensure that all proteins are mutually orthogonal. Results from fluorescence microscopy indicate that all TALEs were efficiently expressed, as determined by presence of mCherry positive cells (Appendix II).

Furthermore, the TALEs efficiently activated gene expression from promoters containing their cognate binding site and not those targeted by other TALEs in the set, indicating that our synthetic TALEs function in a mutually orthogonal manner (Figure 3.5A). Flow cytometry was performed to quantify TALE-activated CFP expression from each promoter. Activity was measured as the total CFP signal of mCherry positive cells. As a control, an off-target TALE was co-transfected with each TALE reporter plasmid and the level of activation for each synthetic TALE was calculated relative to this off-target control. These results confirmed our fluorescence microscopy findings, with synthetic TALEs demonstrating a 10-fold to 102-fold induction of the CFP reporter with no significant signal observed for off-target binding sites (Figure 3.5B, Appendix II).

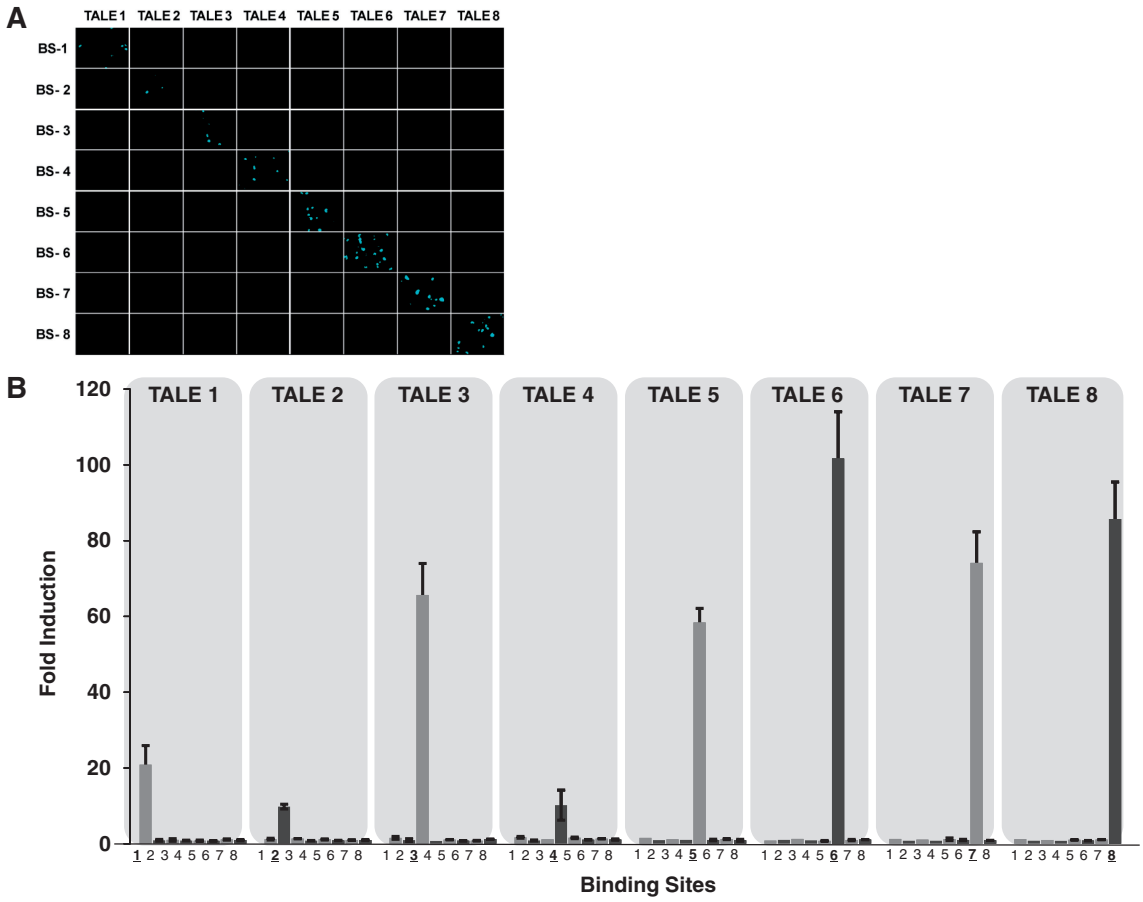


Figure 3.5 Functional characterization of TALE activators

(A) Fluorescence microscopy images of TALE-induced CFP reporter expression. Each column of the 8x8 matrix represents U2-OS cells co-transfected with a synthetic TALE and reporter constructs for each 20-mer binding site (BS). The CFP signal is only visible along the diagonal of the matrix, indicating that the TALEs described here function in a mutually orthogonal manner. (B) Bar graphs representing mutually orthogonal TALE activity as determined by flow cytometry. The fold induction of CFP expression, as calculated relative to an off-target control TALE, displays values ranging from approximately 10-fold to 100-fold for cognate target sites, and demonstrates the functionality and mutual orthogonality of these TALEs.

A

Binding Site	Target Sequence
TALE 8	TATACTATCCAATCCAACTT
TALE 8 + HD-1	TATACTATCCCATCCAACTT
TALE 8 + HD-2	TATACTCTCCAATACAACTT
TALE 8 + HD-3	TATAATATACAATCCCACTT
TALE 8 + HD-4	TATCATATCCAATCCCCCTT
TALE 8 + HD-5	TACCATATCCCATCCAACTC
TALE 8 + HD-7	TCTCATATAAAATAAAACTT
TALE 8 + HD-3 (5`)	TCCCCTATCCAATCCAACTT
TALE 8 + HD-3 (M)	TATACTATCCACCACAACTT
TALE 8 + HD-3 (3`)	TATACTATCCAATCCAAACC

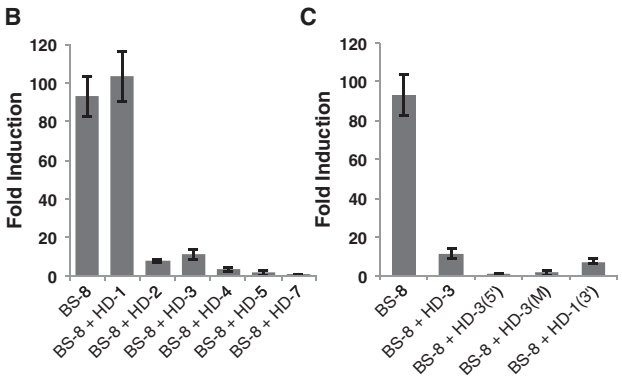


Figure 3.6. Effect of binding site mutations on TALE-mediated transcriptional activation

(A) TALE8 activity in the presence of an increasing number of uniformly distributed binding site mismatches. BS-8 is the corresponding binding site for TALE8 with additional binding sites tested at a hamming distance (HD) of 1 bp to 7 bp from BS-8 (HD-1 to HD-7). The ability of TALE8 to activate CFP expression from each binding site reporter was measured by flow cytometry relative to TALE5 as an off-target control. The presence of two or more mismatches in the binding site significantly decreases the ability of TALE8 to activate gene expression, with binding sites at a hamming distance of more than 3 bp displaying no reporter activity. (B) Effect of binding site mismatch position on TALE activation. The ability of TALE8 to activate gene expression from binding sites with a hamming distance of 3 bp was tested with the position of the mismatches either uniformly distributed, HD-3, localized to the 5'-end of the binding site, HD-3(5'), to the middle of the binding site, HD-3(M), or to the end of the binding site, HD-3(3'). (C) Tested DNA binding sequences. Underlined nucleotides represent mismatches with respect to BS-8.

Synthetic TALEs do not activate transcription of a set of off-target endogenous genes

To investigate the orthogonality of our TALEs to potential genomic promoter binding sites, we began by assessing the effect of target site mismatches on the ability of TALEs to bind a given 20-mer. It has previously been shown that TALE activity generally decreases with the number of mutations in its target site (Kay, Hahn et al. 2009; Romer, Strauss et al. 2009; Scholze and Boch 2010; Miller, Tan et al. 2011; Morbitzer, Elsaesser et al. 2011). However, as positional and contextual effects of these mutations have also been reported, it is important to analyze the specific effect of mutations in the

context of our TALEs that have a different protein architecture and bind to longer DNA sequences (20 bp) than those previously studied. TALE8 was chosen as a representative protein and reporter constructs were generated with 20-mers at a hamming distance of 1 to 7 from TALE8's on-target binding site. To avoid potential position-specific bias, mismatches were distributed evenly throughout the binding sites (Figure 3.6A). TALE8 was co-transfected with each reporter construct, and reporter expression was assayed by fluorescence microscopy and flow cytometry with TALE5 serving as an off-target control. Expression from reporter constructs was observed to decrease with the hamming distance and 20-mers at a hamming distance of 3bp from the on-target site exhibited output signal that was one tenth of the full signal, and 20-mers at a hamming distance of 4 bp or more from the on-target site exhibited an output signal comparable to background (Figure 3.6B).

We next sought to ascertain the influence of mismatch position on protein function. Three additional reporter plasmids were generated for TALE8 with a hamming distance of 3 bp. The positions of these mismatches were localized to either the 5'-end, the 3'-end, or the center of the target site (Figure 3.6A). Our results illustrate that mismatches in the 5'-end and center of the target site abolish TALE activated expression, while mismatches in the 3'-end appear to have less of an impact, more closely resembling mismatches uniformly distributed throughout the binding site (Figure 3.6C). These results indicate that the location of mismatches should be considered when designing orthogonal TALEs. Within the 2kb promoter regoins, the longest matching endogenous sequences to our designed 8 TALEs were at most 14 bp long and these off-target sequences had 4 or more mismatches in positions 14 to 20. Thus, our constructed TALEs satisfy the

combined constraints set by number and position of mismatches in Figure 3.6 and Appendix II.

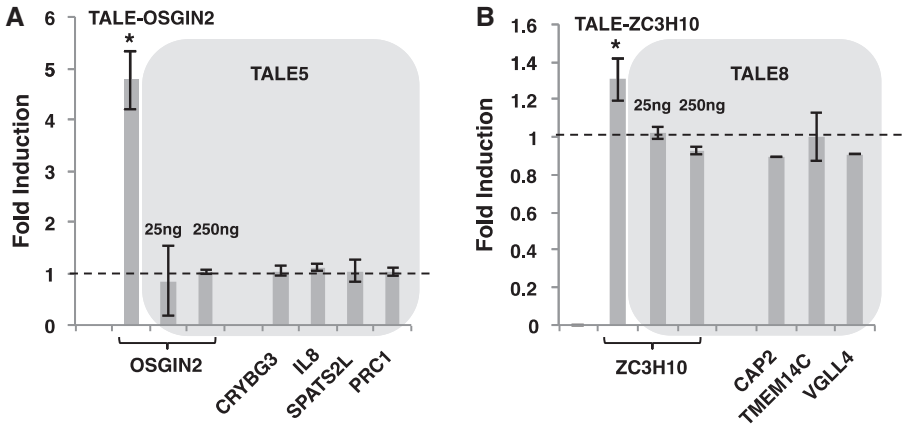


Figure 3.7. Characterization of TALE-mediated off-target endogenous gene activation *in vivo*

Fold change in mRNA levels of potential target genes following TALE expression. mRNA levels of the most likely target genes of TALE5 and TALE8 were measured by qPCR 48 h post-transfection with the corresponding TALE construct and plotted as fold change over mock-transfected cells. TALE-OSGIN2 and TALE-ZC3H10 are the positive control TALEs predicted to activate the two closest off-target genes of TALE5 and TALE8 respectively. (A) A 4.8-fold induction of nearest target gene OSGIN2 by the positive control TALE-OSGIN2, and no significant change in mRNA levels of OSGIN2 and the other four nearest target genes of TALE5 is observed in response to TALE5. The 10x higher concentration (250ng) of TALE5 also shows no significant induction in mRNA levels of its off-target gene OSGIN2. (B) The positive control TALE-ZC3H10 leads to a modest but signficant induction of nearest target gene (ZC3H10) of TALE8. There is no significant change in mRNA levels of the four nearest target genes of TALE8 in response to TALE8 expression. * indicates $P < 0.03$.

To more directly characterize the orthogonality of our synthetic TALEs to endogenous promoter regions *in vivo*, we measured mRNA expression levels from the most likely predicted target genes following transfection with two representative TALEs, TALE5 and TALE8. The nearest predicted off-target sequence for TALE5 was in the promoter of oxidative stress induced growth inhibitor family member 2 (OSGIN2), and

for TALE8 the nearest predicted off-target sequence was in the promoter of zinc-finger CCCH-type containing 10 (ZC3H10). The targets of each TALE chosen for analysis were determined based on the presence of the closest off-target binding site, having a minimum number of mismatches, in the 500 bp region upstream of the transcription start site. As a positive control we designed two TALEs, TALE-OSGIN2 and TALE-ZC3H10, that are predicted to effectively bind in the 500 bp upstream promoter regions of OSGIN2 and ZC3H10, respectively. Off-target sequences for TALEs 5 and 8 and target sequences for TALE-OSGIN2 and TALE-ZC3H10 are listed in Appendix II. All TALEs were transfected in U-2OS cells and the fold change in mRNA level relative to a mock-transfected control was measured at 48 hours post-transfection by qPCR (Figure 3.7).

Results from qPCR demonstrate that while our positive control, TALE-OSGIN2, is capable of inducing OSGIN2 mRNA expression by 4.8-fold, no significant induction is observed following transfection with TALE5 (Figure 3.7A). Similarly, transfection with TALE-ZC3H10 leads to a significant induction of targeted ZC3H10 mRNA, while no significant induction is observed following transfection with TALE8 (Figure 3.7B). In order to ensure that an adequate amount of TALE transcription factor was expressed in cells, we analyzed the fold induction in mRNA expression of off-target genes from TALE5 and TALE8 with 10x higher amount of TALE expression plasmids (Figure 3.7). We observe no significant induction of off-target genes even in the presence of the higher concentration of TALE expression plasmid. To further investigate the orthogonality of our synthetic TALEs, we assayed mRNA expression of the next 4 nearest predicted target genes of TALE5 and the next 3 nearest predicted target genes of TALE8 (Figure 3.7). In all cases, no significant induction of potential target genes was seen relative to mock-

transfected controls, providing further evidence for the orthogonality of these TALEs relative to human promoter regions.

Construction and characterization of TALE repressors

Next, we designed and tested TALE repressor proteins composed of our orthogonal TALE DNA binding domains. We generated TALE repressors TALE5R and TALE8R, by replacing the VP64 activation domain with the KRAB transcriptional repression domain in TALE5 and TALE8 constructs, respectively (Figure 3.8A). The ability of these TALEs to repress transcription was tested by co-transfecting them with CMV-driven CFP expression vectors containing the cognate TALE binding site located on the transcriptional start site of the CMV promoter. TALE repressors efficiently repressed CFP expression from 36 – 97 fold compared to off-target TALE controls (Figure 3.8).

Finally, we demonstrated the ability to tightly repress gene expression to near background levels by combining the TALE repressors with shRNAs targeting the same transcripts. We designed expression constructs that co-express a TALE repressor, an shRNA, and the DsRed fluorescent protein from the same promoter. The shRNAs were generated in the miR30 context and embedded within the SV40 intron in the DsRED red fluorescence protein gene (Stegmeier, Hu et al. 2005; Leisner, Bleris et al. 2010). We used the shRNAs, "FF4" and "FF6," previously designed to target the Firefly Luciferase gene as they are commonly used as off-target negative control shRNAs and are reported to be orthogonal to endogenous transcripts (Leisner, Bleris et al. 2010). When co-expressed, the TALE and shRNA combination repressed CFP expression from 740 –

4853 fold. Of note, the level repression mediated by the TALE repressors alone was at least 5 fold higher than that of the shRNAs expressed alone.

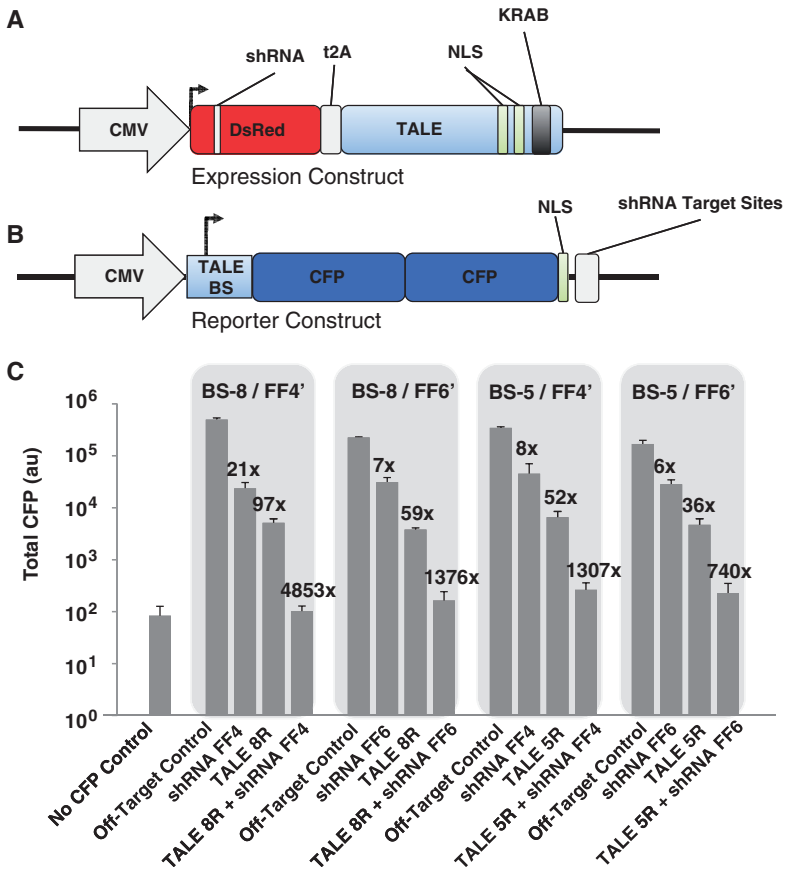


Figure 3.8. Schematics and characterization of TALE repressor-shRNA constructs

(A) The VP64 activation domain of the TALE activators was replaced with the KRAB repression domain and the resulting TALE repressor coding region was cloned into a mammalian expression vector together with a self cleaving DsRed:t2A fluorescent protein. Synthetic shRNAs are expressed from an intron in the DsRed gene. (B) Reporter constructs were generated by cloning a 20 bp TALE target sequence into the transcription start site of the CMV promoter. On binding its recognition site in the promoter, the TALE represses the constitutive expression of the downstream CFP protein. The reporter construct also contains 3 tandem copies of the cognate shRNA recognition sequence in the 3' UTR, which when recognized by the target shRNA leads to degradation of the CFP transcript. (C) TALE repressors, TALE5R and TALE8R were combined with shRNAs, FF4 and FF6, to repress CFP expression from reporter constructs carrying cognate TALE and shRNA recognition sites. Repressions ranged 6x in the case of shRNA alone to approx 4800x in the case of shRNA+TALE repressor combination.

DISCUSSION

Robust synthetic networks would enable the ability to sense a wide variety of cellular cues and respond in a desired fashion to modulate cell behavior, but so far efforts to design these networks have been limited by the reliance on a small set of commonly used gene regulatory components. A large set of mutually orthogonal and modular regulatory components would be a useful tool for generating such networks. Additionally, using components for which interference with the host cell's machinery is minimized would help to reduce the chance of unwanted cellular behaviors and system failures.

TALE transcription factors present a powerful tool with many potential applications including use as a set of reliable gene regulatory components for synthetic gene circuits. However, their utility is limited by degenerate binding and the strong potential for off-target effects (Boch, Scholze et al. 2009; Christian, Cermak et al. 2010; Morbitzer, Romer et al. 2010; Cermak, Doyle et al. 2011; Miller, Tan et al. 2011). While recent work has demonstrated the ability of designer TALE activators to turn on expression of desired genes, they have not been optimized to minimize off-target effects and likely activate the expression of genes other than those intended (Appendix II) (Miller, Tan et al. 2011). Here we present a novel and general method to design TALE DNA binding domains with cognate binding sites orthogonal to a given set of sequences. We create a set of synthetic TALE activators and repressors that specifically recognize and act upon 20 bp binding sites that are at least 3 nucleotide mismatches away from 20 bp sequences contained in all putative human promoter regions. Applying our algorithmic approach to find TALEs that are specific to a given endogenous gene promoter should be relatively less computationally intensive as the search space for such TALEs is very

small compared to the exponentially large search space for TALEs orthogonal to every human promoter. Starting from the set of all possible TALEs that can bind on a given promoter region, the heuristics presented here based on reverse triangle inequality property of hamming distance can be applied to efficiently screen for TALEs that are orthogonal by a given number of base pairs to the rest of the promoters in the genome.

Our synthetic TALE activators displayed high activation of on-target reporters with levels of activation ranging from 10-fold up to 102-fold and are mutually orthogonal. These activation levels are similar to other recently reported TALEs designed to function in mammalian cells, although we employ a different promoter architecture for TALE expression (Miller, Tan et al. 2011). We further characterized the effects of binding site mismatches on TALE orthogonality by selecting a single TALE and generating synthetic target sites containing between 1 and 7 evenly-distributed mismatches. We found that the activation dropped off quickly with an increase in hamming distance – indicating the minimum hamming distance for orthogonality of our TALEs recognizing 20 bp falls in the range of 3-4 bp.

We also found that the distribution of mismatches in the binding site affects TALE protein activity. Testing 20 bp TALE binding sites with sets of three mismatches located at either the 5' end of the binding site, the 3' end of the binding site, the middle of the binding site, or distributed uniformly throughout the 20-mer, we observed that 3 bp mismatches are able to abolish TALE activation when these mutations are introduced at either the 5'end or in the middle of TALE binding site (Figure 3.6C). Three consecutive mutations introduced at the 3'end of binding site show low off-target activity, about one tenth of the full factor, as did the three mutations distributed throughout the binding site.

These results suggest that for binding sites with a 3 bp hamming distance the position of the mutations should be considered.

With these results in mind, we compared the set of 180 computationally derived orthogonal TALE binding sites to all possible 20-mers in 2000 bp upstream promoter regions of the human genome. We found that for genomic sites predicted to be the most likely targets for our synthetic TALEs, the longest region with perfect complementarity from the 5'end was less than 14 bp long for the majority of our synthetic target sites. Furthermore, within this small subset of target sites possessing stretches of sequence complementarity, 4 or more mutations are typically found between positions 13 bp – 20 bp, suggesting that likelihood of a synthetic, orthogonal TALE efficiently binding to a genomic promoter site is extremely low (Appendix II).

To provide further functional evidence for the orthogonality of our synthetic TALEs to genomic promoter regions *in vivo*, we measured mRNA expression levels from the 9 most likely target genes following transfection with two representative TALEs. All potential target genes displayed no increase in mRNA expression levels relative to control, while TALEs designed to specifically target two of those same genes were able to induce mRNA expression up to 4.8-fold. While we cannot rule out the activation of other potential off-target genes by our TALEs, nor the activation of genes by TALE binding to distant enhancer regions outside of the 2kb promoter regions, these results, combined with data detailing the effect of target site mismatches and bioinformatics approaches, provide evidence supporting the ability of our TALEs to function orthogonally to the human promoter regions.

Next, we designed TALE repressors by replacing the VP64 activation domain in the 3' constant back region with the KRAB repressor domains. We assayed two synthetic TALE repressors made from our orthogonal TALE DNA binding domains, along with two synthetic shRNAs, and demonstrate that TALE repressors can provide strong transcriptional repression. The TALE-mediated gene repression was more potent than that accomplished by the two shRNAs tested using the same assay. Double repression of a target gene by the LacI transcriptional repressor and an shRNA was previously reported to be capable of tightly controlling transgene expression (Deans, Cantor et al. 2007). We show that such regulation is also possible by combining TALE repressors and shRNAs. Combined repression reduced the expression level of target protein to near background levels. As TALEs are highly programmable compared to LacI this result allows for the generation of a set of tightly repressed gene modules and opens the possibility of tightly regulating endogenous target genes. TALE repressors have been shown to be a powerful tool for regulating the expression of genes in yeast and plants (Blount, Weenink et al. 2012; Mahfouz, Li et al. 2012). Our results demonstrate that TALE repressors can work efficiently in mammalian cells as well.

Finally, it is worth noting that our proposed algorithm can easily accommodate additional constraints. For example it can be readily adapted to identify orthogonal sequences of different lengths and for different sequences including the genomes of other organisms. It could also be modified to find TALEs that have larger hamming distances to especially critical promoter regions. In addition to addressing the problem of generating synthetic circuit components with minimal effects on endogenous genes, the methods that we employ to generate TALEs are general and can be applied to any system

requiring specific DNA binding domains. Other potential applications of orthogonal TALE DNA binding domains include TALE nucleases, TALE recombinases or TALE-based DNA methylases, and TALE transcriptional activators and repressors that specifically target endogenous genes. The computational approach and transcription factors presented here provide important tools and methods for the precise engineering of biological systems.

ACKNOWLEDGMENTS

The authors wish to acknowledge all members of the Silver lab for helpful comments and discussion. The authors would also like to acknowledge F. Leinert, Y. P. Hung, D. B. Thompson, and L. E. Carey for carefully reading the manuscript. This work was supported by funds from NIH 1F32 CA154195-01 to TZA, the Swiss National Science Foundation (SNSF) fellowship for prospective researchers to AG, grants from DARPA and NIH to PAS, and funds from the Wyss Institute to PAS and JJL.

REFERENCES

An, W. and J. W. Chin (2009). "Synthesis of orthogonal transcription-translation networks."

Proceedings of the National Academy of Sciences of the United States of America **106**(21): 8477-8482.

Andrianantoandro, E., S. Basu, et al. (2006). "Synthetic biology: new engineering rules for an emerging discipline." Mol Syst Biol **2**: 2006 0028.

Barrett, O. P. and J. W. Chin (2010). "Evolved orthogonal ribosome purification for in vitro characterization." Nucleic acids research **38**(8): 2682-2691.

Blount, B. A., T. Weenink, et al. (2012). "Rational diversification of a promoter providing fine-tuned expression and orthogonal regulation for synthetic biology." PloS one **7**(3): e33279.

Boch, J. and U. Bonas (2010). "Xanthomonas AvrBs3 family-type III effectors: discovery and function." Annu Rev Phytopathol **48**: 419-436.

Boch, J., H. Scholze, et al. (2009). "Breaking the code of DNA binding specificity of TAL-type III effectors." Science **326**(5959): 1509-1512.

Carninci, P., A. Sandelin, et al. (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nat Genet **38**(6): 626-635.

Cermak, T., E. L. Doyle, et al. (2011). "Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting." Nucleic Acids Res.

Cermak, T., E. L. Doyle, et al. (2011). "Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting." Nucleic acids research **39**(12): e82.

Christian, M., T. Cermak, et al. (2010). "Targeting DNA double-strand breaks with TAL effector nucleases." Genetics **186**(2): 757-761.

Deans, T. L., C. R. Cantor, et al. (2007). "A tunable genetic switch based on RNAi and repressor proteins for regulating gene expression in mammalian cells." Cell **130**(2): 363-372.

Deng, D., C. Yan, et al. (2012). "Structural basis for sequence-specific recognition of DNA by TAL effectors." Science **335**(6069): 720-723.

Geissler, R., H. Scholze, et al. (2011). "Transcriptional activators of human genes with programmable DNA-specificity." PloS one **6**(5): e19509.

Haynes, K. A. and P. A. Silver (2009). "Eukaryotic systems broaden the scope of synthetic biology." J Cell Biol **187**(5): 589-596.

Herbers, K., J. Conrads-Strauch, et al. (1992). "Race-specificity of plant resistance to bacterial spot disease determined by repetitive motifs in a bacterial avirulence protein." Nature **356**(6365): 172-174.

Kay, S., S. Hahn, et al. (2009). "Detailed analysis of the DNA recognition motifs of the Xanthomonas type III effectors AvrBs3 and AvrBs3Deltarep16." The Plant journal : for cell and molecular biology **59**(6): 859-871.

Khalil, A. S. and J. J. Collins (2010). "Synthetic biology: applications come of age." Nat Rev Genet **11**(5): 367-379.

Knight, T. (2003). "Idempotent Vector Design for Standard Assembly of Biobricks." DSpace. MIT Artificial Intelligence Laboratory;MIT Synthetic Biology Working Group.

Koudritsky, M. and E. Domany (2008). "Positional distribution of human transcription factor binding sites." Nucleic Acids Res **36**(21): 6795-6805.

Lanctot, J. K. L., Ming; Ma, Bin; Wang, Shaojiu; Zhang, Louxin (2003). "Distinguishing String Selection Problems." Informatoin and Computation **185**(1): 41-55.

Leisner, M., L. Bleris, et al. (2010). "Rationally designed logic integration of regulatory signals in mammalian cells." Nature nanotechnology **5**(9): 666-670.

Li, T., S. Huang, et al. (2011). "Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes." Nucleic acids research **39**(14): 6315-6325.

Li, T., S. Huang, et al. (2011). "Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes." Nucleic Acids Res.

Lohmueller, J. J., T. Z. Armel, et al. (2012). "A tunable zinc finger-based framework for Boolean logic computation in mammalian cells." Nucleic acids research.

Lu, T. K., A. S. Khalil, et al. (2009). "Next-generation synthetic gene networks." Nature biotechnology **27**(12): 1139-1150.

MacIsaac, K. D., K. A. Lo, et al. (2010). "A quantitative model of transcriptional regulation reveals the influence of binding location on expression." PLoS Comput Biol **6**(4): e1000773.

Mahfouz, M. M., L. Li, et al. (2012). "Targeted transcriptional repression using a chimeric TALE-SRDX repressor protein." Plant molecular biology **78**(3): 311-321.

Mak, A. N., P. Bradley, et al. (2012). "The crystal structure of TAL effector PthXo1 bound to its DNA target." Science **335**(6069): 716-719.

Miller, J. C., S. Tan, et al. (2011). "A TALE nuclease architecture for efficient genome editing." Nat Biotechnol **29**(2): 143-148.

Morbitzer, R., J. Elsaesser, et al. (2011). "Assembly of custom TALE-type DNA binding domains by modular cloning." Nucleic Acids Res.

Morbitzer, R., J. Elsaesser, et al. (2011). "Assembly of custom TALE-type DNA binding domains by modular cloning." Nucleic acids research **39**(13): 5790-5799.

Morbitzer, R., P. Romer, et al. (2010). "Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors." Proc Natl Acad Sci U S A **107**(50): 21617-21622.

Moscou, M. J. and A. J. Bogdanove (2009). "A simple cipher governs DNA recognition

by TAL effectors." Science **326**(5959): 1501.

Pfaffl, M. W., G. W. Horgan, et al. (2002). "Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR." Nucleic acids research **30**(9): e36.

Phillips, I. S., Pamela (2006). "A New Biobrick Assembly Strategy Designed for Facile Protein Engineering." DSpace. MIT Artificial Intelligence Laboratory;MIT Synthetic Biology Working Group.

Romer, P., S. Recht, et al. (2009). "A single plant resistance gene promoter engineered to recognize multiple TAL effectors from disparate pathogens." Proc Natl Acad Sci U S A **106**(48): 20526-20531.

Romer, P., T. Strauss, et al. (2009). "Recognition of AvrBs3-like proteins is mediated by specific binding to promoters of matching pepper Bs3 alleles." Plant physiology **150**(4): 1697-1712.

Scholze, H. and J. Boch (2010). "TAL effector-DNA specificity." Virulence **1**(5): 428-432.

Scholze, H. and J. Boch (2011). "TAL effectors are remote controls for gene activation." Curr Opin Microbiol **14**(1): 47-53.

Stegmeier, F., G. Hu, et al. (2005). "A lentiviral microRNA-based system for single-copy polymerase II-regulated RNA interference in mammalian cells." Proceedings of the National Academy of Sciences of the United States of America **102**(37): 13212-13217.

Tabor, J. J., H. M. Salis, et al. (2009). "A synthetic genetic edge detection program." Cell **137**(7): 1272-1281.

Van den Ackerveken, G., E. Marois, et al. (1996). "Recognition of the bacterial avirulence protein AvrBs3 occurs inside the host plant cell." Cell **87**(7): 1307-1316.

Wang, B., R. I. Kitney, et al. (2011). "Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology." Nature communications **2**: 508.

Weber, E., R. Gruetzner, et al. (2011). "Assembly of designer TAL effectors by Golden Gate cloning." PloS one **6**(5): e19722.

Xie, X., J. Lu, et al. (2005). "Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals." Nature **434**(7031): 338-345.

Zhang, F., L. Cong, et al. (2011). "Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription." Nature biotechnology **29**(2): 149-153.

Zhu, W., B. Yang, et al. (1998). "AvrXa10 contains an acidic transcriptional activation domain in the functionally conserved C terminus." Mol Plant Microbe Interact **11**(8): 824-832.

Chapter IV

Chimeric Antigen Receptor-based logic gates for re-targeting T-cell specificity

Jason J. Lohmueller collected all of the data in this chapter with mentorship from Yvonne Chen.

ABSTRACT

The ‘re-targeting’ of a cancer patient’s T-cells to kill tumor cells using chimeric antigen receptors (CARs) is a promising new therapy showing recent clinical success. However, currently these therapies target cancer cells based on single cancer-specific biomarkers, limiting the number of cancers that they can treat and leading to off-target effects. To instead target cancers based on multiple markers we engineered a conceptual platform to perform receptor-level ‘logic’ in T-cells, designing novel CARs comprised of native co-stimulatory and co-inhibitory signaling domains. We first report the successful engineering of a CAR-based OR-gate for which T-cell activation occurs in response to target cells expressing tumor antigens CD19, CD20, or both CD19 and CD20. The system consists of two conventional CARs - one that responds to CD19 and one that responds to CD20. We demonstrate the efficacy of this system in the Jurkat T-cell line based on T-cell activation marker staining and cytokine release assays. We also demonstrate progress toward a NOT gate system that can activate T-cell signaling in response to CD19 but inhibit this signaling by the presence of CD20 on the same cells. We generated putative inhibitory CARs using the cytoplasmic domains of the CD45 and CD300a co-inhibitory receptors. We demonstrate successful engagement by CD20 these receptors in Jurkat cells but no effect on surface staining or specific target cell lysis. However, in follow-up assays in primary T-cells for the CD20-CD8hinge-CD300a receptor we observed a mild, but-significant inhibition of specific cell lysis T-cells. The new targeting specificities enabled by our receptor logic systems could be used to expand the number of cancers treatable by these targeted T-cell therapies and greatly minimize off-tumor effects.

INTRODUCTION

Chimeric antigen receptors (CARs) are artificial T-cell receptors that allow the re-targeting of cytotoxic T-cell activity toward cells expressing a defined surface antigen. CAR-based immunotherapies have recently emerged as promising treatments for cancer and have shown success in trials for leukemias and metastatic melanoma. In one recent trial treating chronic lymphocyte leukemia over 65% of patients displayed an ongoing complete response to therapy after 10 months (Porter, Kalos et al. 2011; Porter, Levine et al. 2011).

As stated in their name, CARs are chimeric molecules consisting of T-cell Receptor and co-receptor cytoplasmic signaling domains, a transmembrane domain, an extracellular linker domain, and finally a targeting domain, most often an scFv. The targeting domain is modular, and by using scFvs for different antigens researchers can easily change the specificity of the CAR and the resulting T-cell targeting. The use of an an scFv for targeting instead of a normal T-cell Receptor (TCR) also allows the CAR to be used in patients without the need for MHC-matching as in the case of other native TCR-based immunotherapies (Sadelain, Riviere et al. 2003).

The engineering of CARs has undergone three generations of development each including modifications to the cytoplasmic portion of the receptors. The first generation contained only the CD3 zeta chain from the TCR in the cytoplasm. While these receptors were capable of specifically activating T-cells in response to antigen, the cells did not survive and expand enough in mice or human patients to be effective. For the 2nd generation of CARs the CD28 co-receptor was added to the cytoplasmic portion of the protein. CD28 signaling is known to induce the cytokine IL-2 leading to prolonged cell-

survival. These receptors had more efficacy *in vivo,* but in the meantime researchers found that adding yet another domain, the 4-1BB domain known to enhance T-cell expansion, in combination with the CD28 domain led to an even better *in vivo* response (Porter, Levine et al. 2011).

While individual CARs have been developed to target many different tumor-associated antigens, currently each therapy only uses one receptor and targets tumors via a single tumor-specific antigen. This single antigen targeting is limiting in many cases as it is often difficult to find such a specific antigen for a given cancer. Even in current therapies many "on-target off-tumor" effects are seen which can lead to sometimes-lethal levels of toxicity (Lamers, Sleijfer et al. 2006; Brentjens, Riviere et al. 2011; Kalos, Levine et al. 2011; Kochenderfer, Dudley et al. 2012).

We look to reduce these "on-target off-tumor effects" and increase the number of cancers targetable by creating T-cell logic gates that target tumors based on the presence or absence of multiple antigens. Natural T-cell signaling is a process that involves many co-stimulatory and inhibitory receptors and so is naturally inclined toward receptor-based logical engineering. One very recent example of this has been demonstrated in mice by Kloss et al. who created an AND gate using a novel pair of CARs. Their system consisted of a 1st GEN CAR with a weakened scFv domain and a novel co-stimulatory CAR with a full strength antibody targeting a second antigen fused to the CD28 and 4-1BB signaling domains. The system showed efficacy *in vitro* and in mouse studies (Kloss, Condomines et al. 2012).

We first sought to engineer a CAR-based OR-gate that would respond to two individual antigens and both antigens together. To engineer this system we used a simple

design of expressing two 2nd GEN CARs for two different antigens in cells and analyzed the cells' behavior in co-incubation assays. Interestingly while the OR-gate system does not add greater specificity to T-cell therapy, recent work by Landsberg et al. on clinical resistance to T-cell therapy pointed to an OR-gate system as a possible strategy to obviate resistance. They found that in response to adoptive immunotherapy targeting a single melanoma antigen, the melanoma cells began to differentiate via a cell defined signaling pathway and so as to longer express the targeted gene. They posit that targeting of the first antigen and a second antigen expected to be on the differentiated cells could stop this differentiation process and eliminate differentiated cells (Landsberg, Kohlmeyer et al. 2012). We also sought to construct an A AND NOT B targeting system. To create this gate we combined a conventional CAR for positive signaling (antigen A) and generated putative inhibitory CARs containing fragments of known inhibitory co-receptors. For these inhibitory CARs we used the cytoplasmic domains of CD45 and CD300a. Both of these molecules contain immunoreceptor tyrosine-based inhibition motifs (ITIMs) known to recruit phosphatases such as SHP-1 and SHP-2 that inhibit T-cell signaling (Hermiston, Zikherman et al. 2009; DeBell, Simhadri et al. 2012). As the extracellular spacings of many inhibitory receptors are thought to be important for effects on TCR signaling we made different versions of these receptors with different extracellular linker domains (James and Vale 2012).

RESULTS AND DISCUSSION

We first set out to generate a CAR-based OR-gate that would activate T-cells when faced with target cells expressing CD19, CD20, or CD19+CD20. We designed the

system to be comprised of two 2nd generation CARs each targeting either CD19 or CD20. The receptors are comprised of the CD3-zeta chain and the CD28 co-stimulatory domain in the cytoplasm, the CD28 transmembrane domain, and the Igg4 linker domain and an scFv targeting either CD19 or CD20 on the cell surface (Figure 4.1A).

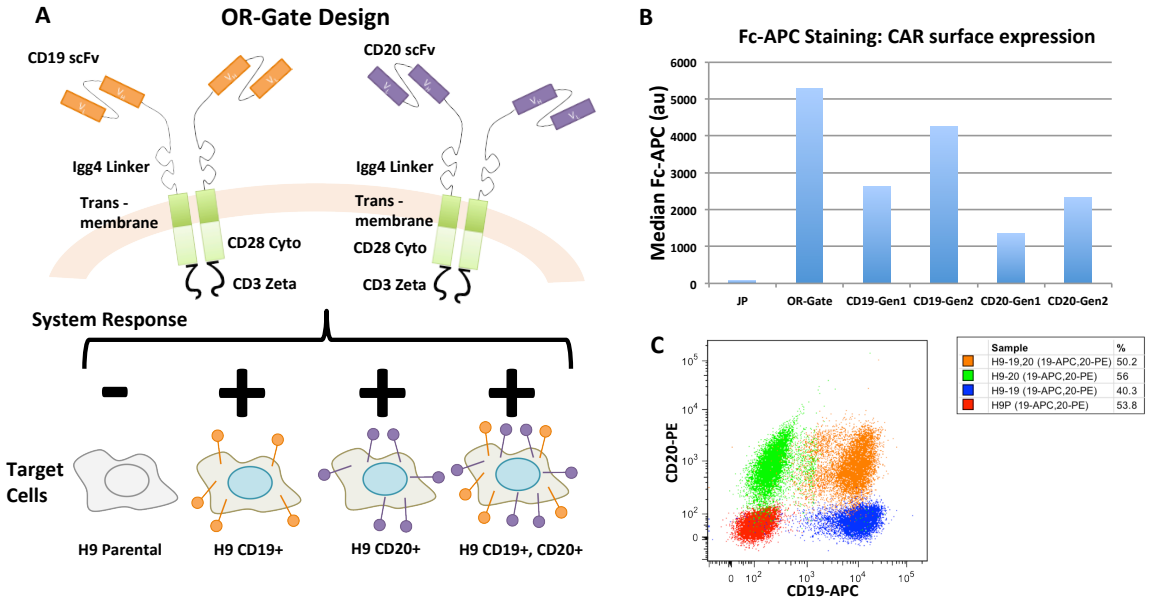


Figure 4.1 OR-Gate design and expression characterization

(A)The 2-CAR OR-Gate system is expected to activate T-cell signaling in response to target H9 cells that are CD19+, CD20+, OR CD19+CD20+. (B) Median Fc-APC staining for Jurkat cells, to report on CAR expression. Fc binds to the Igg4 linker domain thus labeling both CARs. Of note for the OR-Gate it cannot be determined the ratio of both CARs only the total Fc-APC staining level. (C) Staining of the antigens on the target cells by CD19-APC and CD20-PE antibodies.

We generated lentivirus encoding these receptors and created stable Jurkat cell lines encoding single CARs and the two CAR OR-gate system. We then verified expression of the CARs on the cell surface by staining with a dye-conjugated Fc domain and flow cytometry and sorted for stable cell lines based on these markers (Figure 4.1B). To test the system we generated three H9 target cell lines expressing CD19, CD20, or CD19+CD20. As H9 cells don't naturally express either CD19 or CD20 we crated these

lines by transducing them with virus encoding these antigens. We verified surface expression of these molecules and sorted for antigen positive lines again by antibody staining for these markers and flow cytometry (Figure 4.1C). These cells were used as target cells in all co-incubation assays in this chapter.

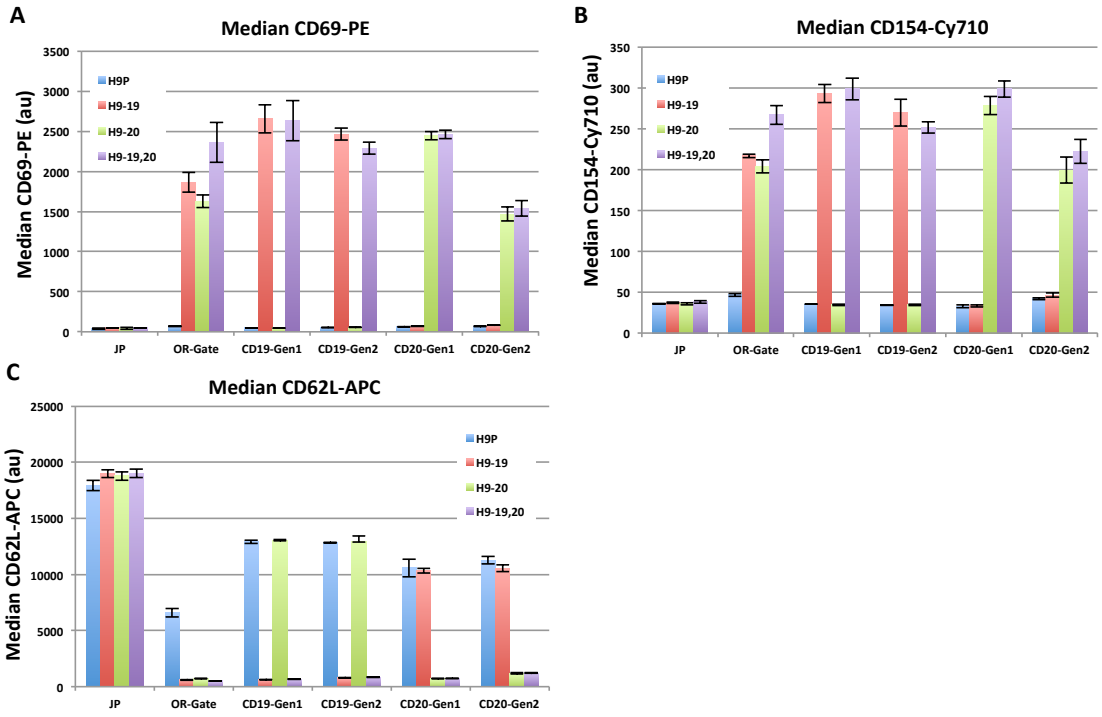


Figure 4.2 OR-Gate T-cell activation marker staining

(A) CD69-PE staining of single CAR and OR-Gate 24hr co-incubations with H9 Target cells +/- std. dev. (n=3). (B) CD154-Cy710 staining of single CAR and OR-Gate 24hr co-incubations with H9 Target cells +/- std. dev. (n=3). (C) CD62L--APC staining of single CAR and OR-Gate 24hr co-incubations with H9 Target cells +/- std. dev. (n=3).

Next, we performed effector T-cell and target cell co-incubation assays to test for antigen specific activation. To distinguish between H9 and Jurkat cell lines, before the assay we stained all H9 target cell lines with the CellTracker dye. We then co-incubated Jurkat parental (JP), single 1st GEN CAR, single 2nd GEN CAR, and OR-gate cell lines with each of the four stained H9 target cell lines (H9P, H9+CD19, H9+CD20 and

H9+CD19+CD20) for 24hrs. Following this co-incubation we stained the cells for T-cell activation markers: CD69, CD62L and CD154, and analyzed expression of these markers by flow cytometry (Figure 4.2A,B,C). Markers CD69 and CD154 are known to be up-regualted by T-cell activation while CD62L is expected to be down-regulated. We found that as expected single 1st GEN and 2nd GEN CARs specifically and efficiently activated markers CD69 and CD154 and down-regulated CD62L in response to target cells. The OR-gate effector cells also functioned as desired, activating CD69 and CD154 and down-regulating CD62L in response to target cells with single antigens or both antigens.

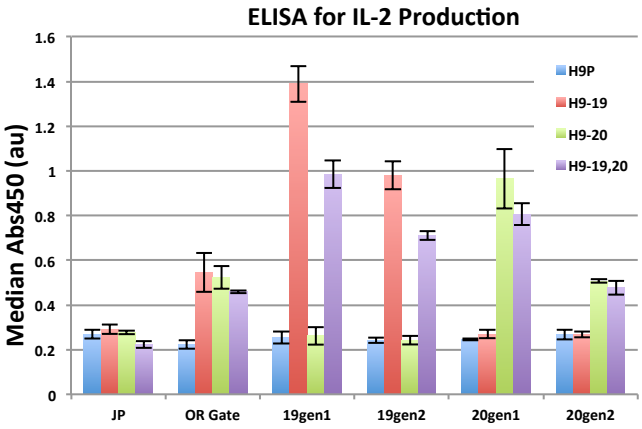


Figure 4.3 OR-Gate IL-2 ELISA assay

IL-2 production of single CAR and OR-Gate 48hr co-incubations with H9 Target cells +/- std. dev. (n=3).

We then tested these same lines for another indicator of T-cell activation, IL-2 production. IL-2 is an important survival signal produced by activated T-cells to assist in their expansion. Co-stimulatory signaling by CD28 is expected to enhance its production and addition of the CD28 to the 2nd GEN CARs is largely for this reason. We again performed co-incubation assays of the target cell lines with the different single CAR and OR-gate effector lines. Following 48hr co-incubuation we performed an ELISA for IL-2

activity. We found that, once again, the single CAR and OR-gate effector cell lines showed signs of being efficiently and specifically activated, producing IL-2 only when co-incubated by the correct target cells. Surprisingly the single 1st GEN CARs produced the highest amounts of IL-2 even though they did not contain the CD28 co-stimulatory domain. OR-gate and CD20 2nd GEN CAR IL-2 inductions were somewhat lower but still significant and specific (Figure 4.3). Interestingly, the IL-2 production and action marker staining did not correlate with the expression levels of the receptors as observed in Figure 4.1A,B,C suggesting that the 1st GEN receptors could simply be more efficient at stimulating T-cell activation.

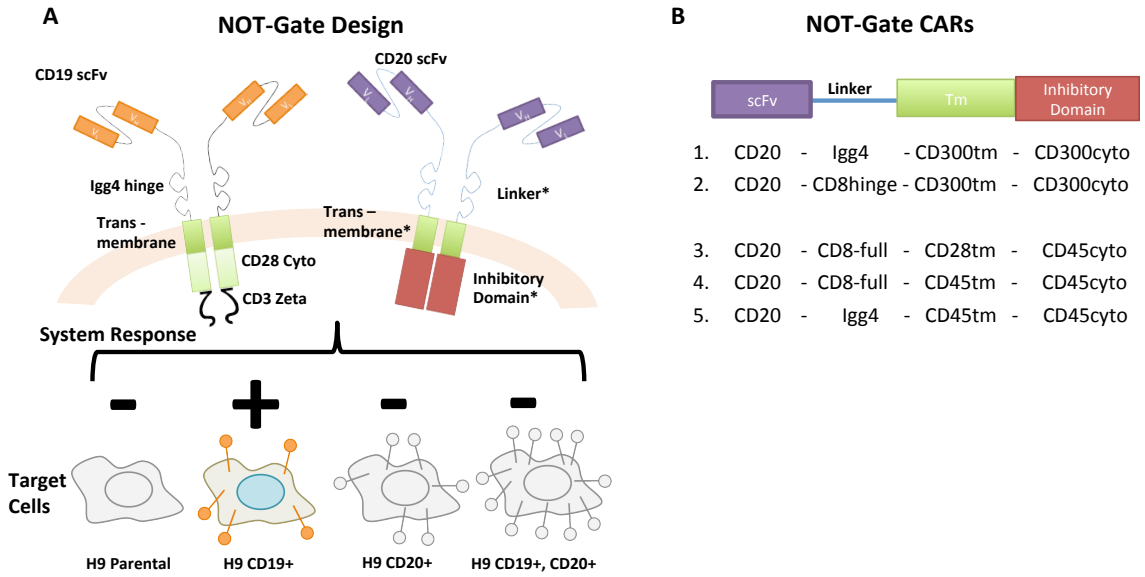


Figure 4.4 NOT-Gate design and receptor variants

(A)The 2-CAR NOT-Gate system is expected to activate T-cell signaling in response to target H9 cells that are CD19+ AND NOT CD20+. (B) The 5 NOT-gate CAR variants are listed combining different linker domains, transmembrane domains (Tm), and inhibitory cytoplasmic domains.

Next, we sought to build inhibitory CARs using the CD45 and CD300a inhibitory signaling domains. As the extracellular dimensions of natural inhibitory signaling

receptors relative to TCR complexes and transmembrane domains (for signaling) have been shown to be important parameters for natural T-cell signaling and inhibition we created inhibitory CARs with differing extracellular linkers and transmembrane domains (Cahir McFarland and Thomas 1995; Choudhuri, Wiseman et al. 2005). We varied the receptor heights by using extracellular linker domains of different lengths. We generated DNA constructs encoding a set of putative inhibitory CARs targeting CD20 and additionally labeled with a co-translationally expressed TagBFP fluorescent reporter as outlined in Figure 4.4A. We then made lentivirus and Jurkat cell lines encoding these receptors alone, or with a conventional CD19 2nd GEN CAR that is tagged with a co-translationally expressed EGFRt surface marker that can be stained by Erbitux antibody (Figure 4.4B). We confirmed expression of these receptors and selected positive lines by staining with Protein-L, a protein that is expected to bind all scFv fragments.

We then performed co-incubation assays using these cells expressing these and the four H9 target target cell lines. Once again we labeled the H9 target cells with CellTracker dye prior to co-incubation so that we could isolate the Jurkat effector cells. After 24hrs we stained the cells for T-cell activation markers CD25, CD69, CD154, and CD62L (Figure 4.5A,B,C; Figure 4.6A,B,C). We found that while the conventional CD19-CAR was capable of activating and repressing the expected T-cell activation markers, the inhibitory receptors, did not inhibit these markers as hoped for the dual+ target cell lines. We did find however, that some of the CD20 inhibitory receptors were capable of mildly inducing some of the markers suggesting target receptor engagement. It is unclear what this mild activation could mean for the T-cell signaling process.

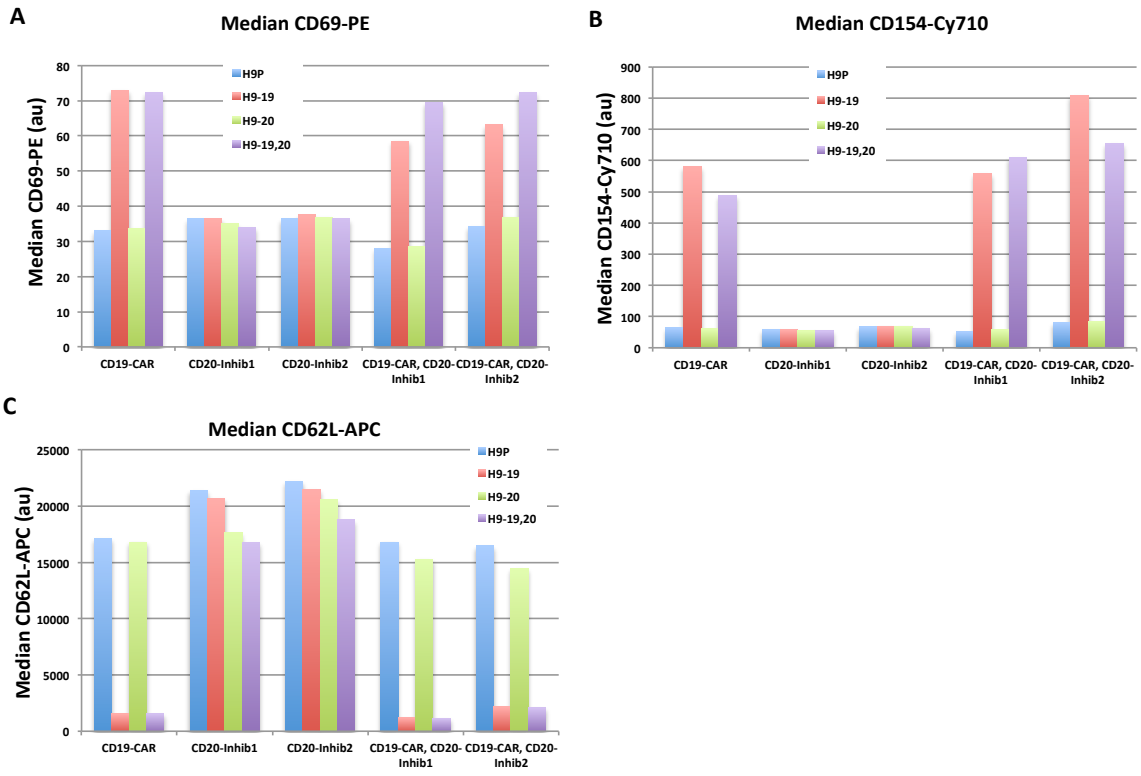


Figure 4.5 NOT-Gate T-cell activation marker staining for inhibitory receptors 1 and 2

(A) CD69-PE staining of single CAR and NOT-Gate 24hr co-incubations with H9 Target cells (n=1). (B) CD154-Cy710 staining of single CAR and NOT-Gate 24hr co-incubations with H9 Target cells (n=1). (C) CD62L--APC staining of single CAR and NOT-Gate 24hr co-incubations with H9 Target cells (n=1).

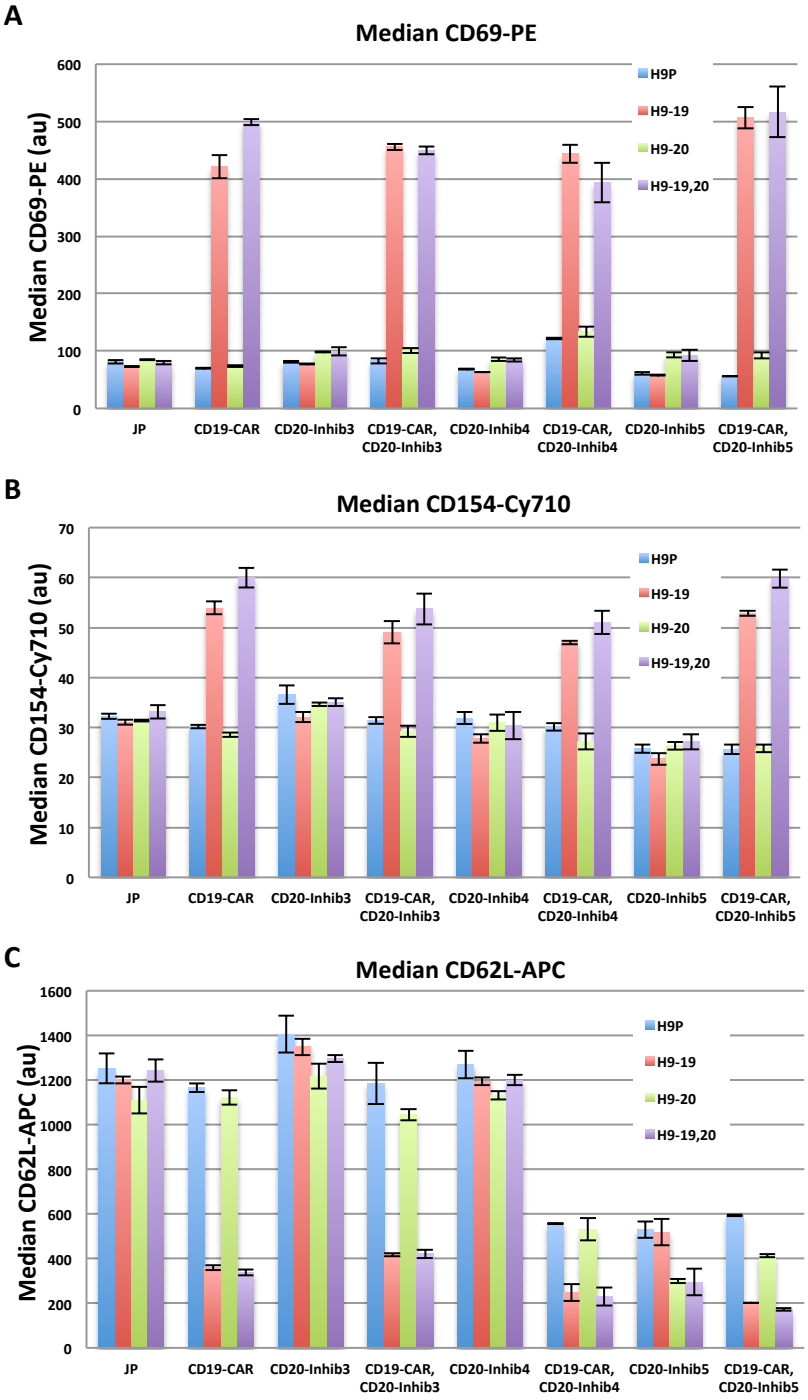


Figure 4.6 NOT-Gate T-cell activation marker staining for inhibitory receptors 3,4 and 5

(A) CD69-PE staining of single CAR and NOT-Gate 24hr co-incubations with H9 Target cells +/- std. dev. (n=3). (B) CD154-Cy710 staining of single CAR and NOT-Gate 24hr co-incubations with H9 Target cells +/- std. dev. (n=3). (C) CD62L--APC staining of single CAR and NOT-Gate 24hr co-incubations with H9 Target cells +/- std. dev. (n=3).

As an additional assay we sought to directly look at T-cell mediated target cell lysis. While Jurkat cells are CD4+ T-cells and are conventionally expected to be 'helper' cells as opposed to conventional CD8+ cytotoxic T-cells, researchers have shown CD4+ cells are also capable of targeted cell lysis. We performed specific lysis co-incubation assays using the NOT gate cell lines (Figure 4.7A,B,C,D,E,F). We found the expected levels of lysis for the conventional CD19-CAR, however no effect in the NOT gate lines, we saw no inhibition of the specific lysis for the CD19+,CD20+ dual positive line compared to the CD19+ line. Unfortunately, these results suggest the inhibitory receptors are non-functional in the Jurkat T-cell line.

As a last ditch effort we transduced harvested CD8+ T-cells with viruses encoding the conventional CD19 CAR and one of the inhibitory CARs, the CD20scFv-Igg4-CD300a inhibitory CAR. We sorted the cells for TagBFP expression and expanded them. We then performed specific lysis assays using these cells (Figure 4.8A,B). We now saw

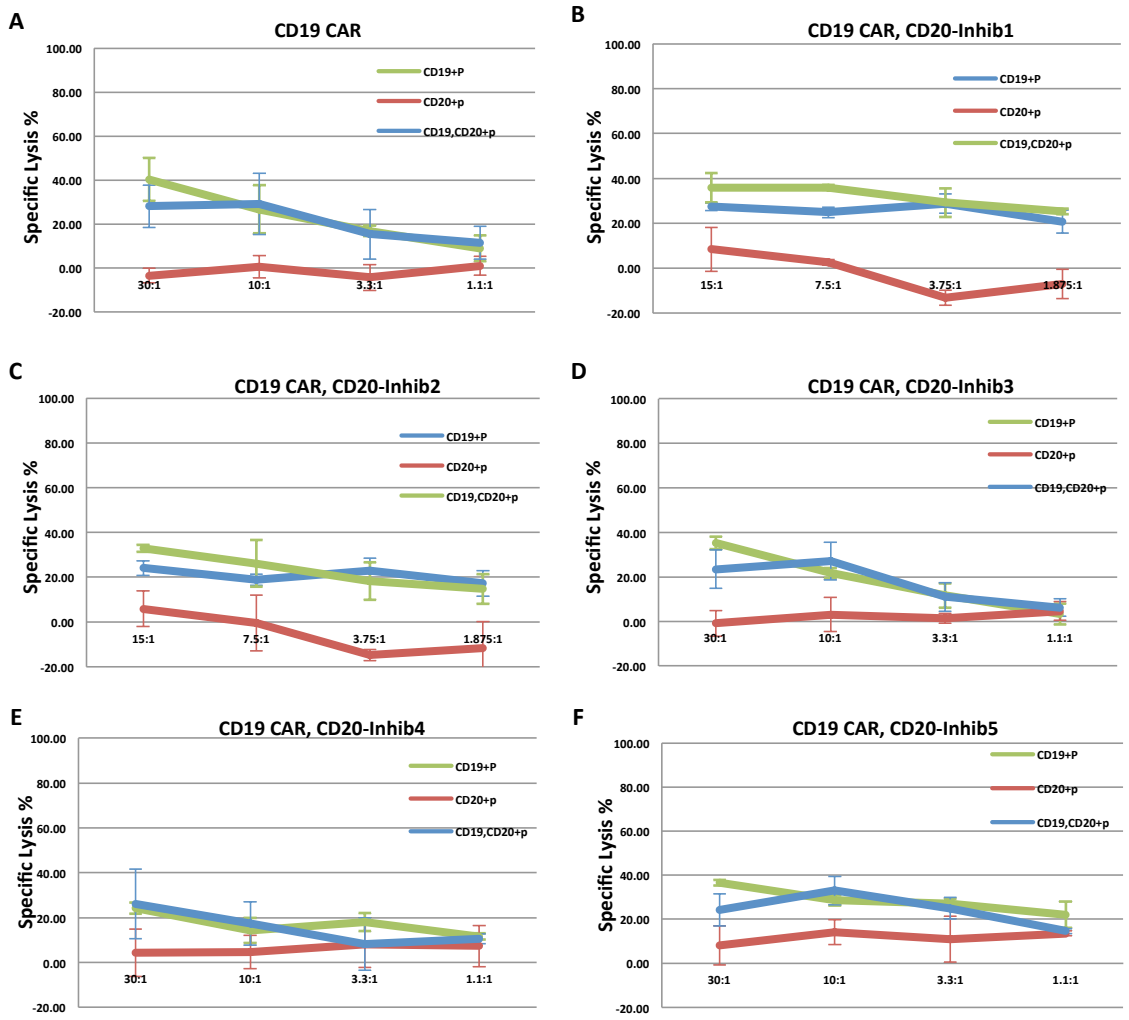


Figure 4.7 Specific lysis assays for Jurkat NOT-Gate cell lines

(A) Specific lysis for the control CD19-CAR alone assayed after 4hrs+/- std. dev. (n=3). (B) Specific lysis for the control CD19-CAR,CD20-Inhib1 NOT-gate assayed after 4hrs+/- std. dev. (n=3). (C) Specific lysis for the control CD19-CAR,CD20-Inhib2 NOT-gate assayed after 4hrs +/- std. dev. (n=3). (D) Specific lysis for the control CD19-CAR,CD20-Inhib3 NOT-gate assayed after 4hrs+/- std. dev. (n=3). (E) Specific lysis for the control CD19-CAR,CD20-Inhib4 NOT-gate assayed after 4hrs. +/- std. dev (n=3). (F) Specific lysis for the control CD19-CAR,CD20-Inhib5 NOT-gate assayed after 4hrs +/- std. dev. (n=3).

a mild but significant inhibition in killing of the CD19+, CD20+ cell line compared to the

CD19+ line for the CD20scFv-Igg4-CD300a NOT gate system. The CD19-CAR-only

system killed both target lines efficiently further suggesting the observed difference with the CD300a inhibitory receptor is a result of the inhibitory receptor. As this difference is

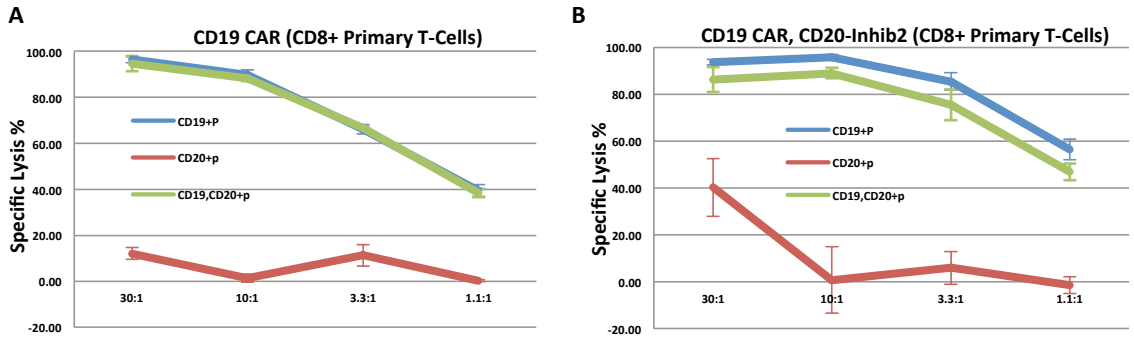


Figure 4.8 Specific lysis assays for Primary NOT-Gate cell lines

(A) Specific lysis for the control CD19-CAR alone assayed after 4hrs +/- std. dev. in primary T-cells (n=3). (B) Specific lysis for the control CD19-CAR,CD20-Inhib2 NOT-gate assayed after 4hrs +/- std. dev. in primary T-cells (n=3).

so mild it is unlikely to lead to any significant physiological effect in its current stage however, it is possible that this system could be improved to show more robust inhibitory behavior. One potential approach is to vary the ratio of the positive and negative receptors. Another potential approach would be to add more of the known ITIM domains in the CD300a receptor to the cytoplasmic domain to try to increase this inhibition on a per-receptor basis. Finally, the apparent success with this inhibitory receptor that showed no effect in Jurkat cells suggests that Jurkat cell assays do not reflect the real activities of the primary cells and that the other receptors might be more effective in the primary cells.

CONCLUSION

The targeting of cancers based on multiple antigens would grant much-needed additional specificity to cancer immunotherapies. We demonstrate here a functional CAR-based OR-gate in Jurkat T-cells. OR-gate cells show antigen targeted expression of

both T-cell activation surface markers and IL-2 cytokine production in an OR-gate manner. The next steps will be to show the efficacy of these cells in primary T-cells, and in tumor-targeting mouse studies. As suggested by Landsberg et al. this OR-Gate system could be used to block cancer resistance to immunotherapy. We next sought to create an A AND NOT B system, and generated several putative NOT receptors based on the CD45 and CD300a inhibitory receptors. All receptors were expressed and appeared to engage target antigen, however they had no effect on T-cell signaling in Jurkat cells. We then tried one of the CD300a recpetors in primary T-cells and found that the receptor had a mild effect on specific cell lysis. While not useful in its current state, this inhibitory receptor could be modified to increase its inhibitory activities either through higher relative expression to the positive signaling CAR or an increase in the ITIM domains on the receptor.

MATERIALS AND METHODS

DNA assembly and purification

To generate CAR constructs we synthesized codon-optimized DNAs encoding the CD19 and CD20 scFvs, the Igg4 linker domain, the CD3zeta chain, the CD28-cytoplasmic domain, and the CD300a cytoplasmic domain. We PCR'd the CD45 cytoplasmic domain from a cDNA purchased from Origene technologies (Rockville, MD) and the CD8 linker domain from a cDNA plasmid purchased from Origene technologies (Rockville, MD). We created the Igg4-hinge and CD8-hinge domains by PCR from the longer linker forms. Using these parts we then assembled the CAR expression constructs using Gibson assembly, expressing them from the EF1alpha

promoter in the epHIV7 lentiviral vector. Plasmids were grown up in E. coli and Maxi-prepped using Endotoxin Free Kits (Qiagen; Hilden, Germany).

Virus production and transduction

We plated HEK293T in 10cm^2 plates 24hrs prior to Calphos transfection with the expression construct and 3 packaging plasmids using the CalPhos Transfection Kit (Clontech Laboratories, Inc.). 18hrs following transfection media was changed adding 5M Sodium Butyrate. Viral supernatants were then harvested every 24hrs for 3 days. These harvests were centrifuged and filtered to remove cell debris, and then concentrated with PEG overnight. The following day these harvests were further concentrated via ultracentrifugation, and frozen at -80C° until further use. For transduction we thawed virus and added it to 5*10^4 cells with 4ug/ml polybrene.

Cell culture and maintenance

We cultured Jurkat T-cells clone E6-1 (ATCC CAT# TIB-152) and their derivatives and H9 Cells (ATCC CAT# HTB-176) and their derivatives in suspension in RPMI media supplemented with 10% FBS and 1% PenStrep (Life Technologies; Carlsbad, CA). For cell maintenance passaged cells every 3-5 days by diluting them 1:10. On the day of harvest we isolated T-lymphocytes from whole cord blood by standard Phycol preparation. We then stimulated and isolated CD8+ T-cells using the CD8+ T-cell isolation Kit from Miltenyi Biotec (Cambridge, MA). We expanded the cells by stimulating them with IL-15 and IL-12 cytokines every 2 days.

Co-incubation and surface staining assay

For surface staining assays, we co-incubated 5*10^5 Jurkat effector cells with an equal number of an H9 target line in 1 mL of supplemented RPMI media in 24well plates for 24hrs. Following co-incbation we moved cells to 96-well plates and washed them with PBS 2x. We then stained them with the antibodies at the appropriate dilutions for 30 minutes on ice. Finally, we washed them again with PBS 2x and assayed via flow cytometry. Antibodies and there vendors are listed CD19-APC (Clone LT19, Miltenyi Biotec), CD20-PE (Clone LT20, Miltenyi Biotec), CD62L-APC (Clone DREG-56 eBioscience), CD69 (Clone FN50, BioLegend), CD154 (Clone 24-31 eBioscience) Human IgG Fc-APC ([No clone] eBioscience).

Co-incubation and ELISA assay

We co-incubated 5*10^5 Jurkat effector cells with an equal number of an H9 target line in 1mL of supplemented RPMI media in 24well plates for 48hrs. For the ELISA we isolated the supernatant from the wells and processed them using the IL-2 ELISAMAX Kit from BioLegend (San Diego, CA) precisely following the supplied protocol.

Co-incubation and specific cell lysis assay

Prior to co-incubation we stained H9 Parental cells with CFSE dye (Life Technologies; Carlsbad, CA), and potential target cells with CMTMR (Life Technologies; Carlsbad, CA). We then co-incubated 10^4 H9-Parenal cells and 10^4 of a given target cell line with different ratios of each effector T-cell line being tested in 100ul

of RPMI media in a 96-well plate. (E:T ratios - 30:1, 10:1, 3.3:1, and 1.1:1) Following a 4hr incubation we analyzed samples by flow cytometry and for a given effector line determined the ratio of CMTMR target cells: CFSE H9 Parental cells as a ratio of these cells in MOCK co-incubations to determine specific lysis.

REFERENCES

Brentjens, R. J., I. Riviere, et al. (2011). "Safety and persistence of adoptively transferred autologous CD19-targeted T cells in patients with relapsed or chemotherapy refractory B-cell leukemias." Blood **118**(18): 4817-4828.

Cahir McFarland, E. D. and M. L. Thomas (1995). "CD45 protein-tyrosine phosphatase associates with the WW domain-containing protein, CD45AP, through the transmembrane region." The Journal of biological chemistry **270**(47): 28103-28107.

Choudhuri, K., D. Wiseman, et al. (2005). "T-cell receptor triggering is critically dependent on the dimensions of its peptide-MHC ligand." Nature **436**(7050): 578-582.

DeBell, K. E., V. R. Simhadri, et al. (2012). "Functional requirements for inhibitory signal transmission by the immunomodulatory receptor CD300a." BMC immunology **13**: 23.

Hermiston, M. L., J. Zikherman, et al. (2009). "CD45, CD148, and Lyp/Pep: critical phosphatases regulating Src family kinase signaling networks in immune cells." Immunological reviews **228**(1): 288-311.

James, J. R. and R. D. Vale (2012). "Biophysical mechanism of T-cell receptor triggering in a reconstituted system." Nature **487**(7405): 64-69.

Kalos, M., B. L. Levine, et al. (2011). "T cells with chimeric antigen receptors have potent antitumor effects and can establish memory in patients with advanced leukemia." Science translational medicine **3**(95): 95ra73.

Kloss, C. C., M. Condomines, et al. (2012). "Combinatorial antigen recognition with balanced signaling promotes selective tumor eradication by engineered T cells." Nature biotechnology **31**(1): 71-75.

Kochenderfer, J. N., M. E. Dudley, et al. (2012). "B-cell depletion and remissions of malignancy along with cytokine-associated toxicity in a clinical trial of anti-CD19

chimeric-antigen-receptor-transduced T cells." Blood **119**(12): 2709-2720.

Lamers, C. H., S. Sleijfer, et al. (2006). "Treatment of metastatic renal cell carcinoma with autologous T-lymphocytes genetically retargeted against carbonic anhydrase IX: first clinical experience." Journal of clinical oncology : official journal of the American Society of Clinical Oncology **24**(13): e20-22.

Landsberg, J., J. Kohlmeyer, et al. (2012). "Melanomas resist T-cell therapy through inflammation-induced reversible dedifferentiation." Nature **490**(7420): 412-416.

Porter, D. L., M. Kalos, et al. (2011). "Chimeric Antigen Receptor Therapy for B-cell Malignancies." Journal of Cancer **2**: 331-332.

Porter, D. L., B. L. Levine, et al. (2011). "Chimeric antigen receptor-modified T cells in chronic lymphoid leukemia." The New England journal of medicine **365**(8): 725-733.

Sadelain, M., I. Riviere, et al. (2003). "Targeting tumours with genetically enhanced T lymphocytes." Nature reviews. Cancer **3**(1): 35-45.

Chapter V

CONCLUSION

The future of mammalian synthetic gene networks appears to be very bright. The lists of possible circuit inputs and circuit functionalities are ever-expanding, and recent advances to creating computational frameworks and genome engineering tools will pave the way for building even more sophisticated circuits. In the preceding chapters we described our three primary contributions to the engineering of mammalian circuits including work on a zinc finger computing framework, orthogonal TALE transcriptional regulators, and chimeric antigen receptor-based T-cell targeting. Here are some concluding thoughts on these projects.

The zinc finger systems we developed in Chapter II provide a large set of transcriptional circuit components with varying activities and approaches to building logic circuits. The method of increasing transcription factor (TF) regulator activity through transcription factor dimerization was novel for artificial regulators and perhaps could be applied to other TFs including those controlling endogenous gene expression. The split intein-based AND and NAND gates were also novel and provided a robust means of performing AND-logic. However, the systems are not very scalable as every new AND gate requires a new intein pair. Every 2-hybrid-based AND gate similarly requires two unique interaction domains per AND gate, however more interaction domains exist than split intein pairs. One potential solution to creating a more scalable system could be to split transcription activation and repression domains such that the halves have no activity on their own and must be brought together by DNA binding domains on the DNA. In such a system AND gates would only require new DNA binding domains for each successive AND gate. We also report a new protein induced intein that

essentially represents a 3-input AND gate, and could be of future biotechnological use in Appendix III.

The orthogonal TALE transcription factors that we created are ideal circuit components as they are not expected to bind to any human promoter regions. In addition to the evidence that we provided using qRT-PCR analysis of expected off-target genes, and assays using synthetic off-target promoters, it will be important for these factors to be tested on a genome-wide scale. Recent studies examining off-target effects of zinc finger nucleases suggest that off-target binding by DNA binding domains might not be as predictable as anticipated.

The CAR-based OR-Gate that we designed functioned very well in Jurkat cell line. The next steps will be to try it in primary T-cells and then in an animal model. It would also be interesting apply the system to target a differentiated tumor state, and thus block resistance to therapy. It would also be interesting to compare the activity and efficacy of the 2-CAR OR-gate with that of two different populations of T-cells, each containing only one of the 2-receptors. Such a therapy might be easier to implement than one requiring two receptors expressed on the same cell population, however it may have better functionality. The final result with the CD20-CD8hinge-CD300a CAR-based NOT system was very mildly positive. This result perhaps could be improved by increasing the ratio of negative to positive receptor in the cells, or by increasing the number of ITIMs on the cytoplasmic domain. If these changes don't lead to an increase in inhibition, there are also many other inhibitory co-receptors available to try. In particular the KIR and LIR NK cell receptors are promising as their engagement with target cells leads directly to suppression of cytotoxic signaling. It will also be interesting to see if it is possible to

create novel CARs to suppress immune responses using tolerogenic receptors to combat auto-immune diseases.

Appendix I

Supplemental Information: A tunable zinc finger-based framework for Boolean logic computation in mammalian cells

Jason J. Lohmueller[1,2], Thomas Z. Armel[1] & Pamela A. Silver[1,2]

[1] *Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA*

[2] *Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts 02115, USA*

Reproduced from Lohmueller JJ, Armel TZ, Silver PA. (2012). A tunable zinc finger-based framework for Boolean logic computation in mammalian cells. *Nucleic Acids Res.* Jun;40(11):5180-7.

Jason J. Lohmueller contributed all data for figures in collaboration with Thomas Z. Armel.

Inventory of Supplemental Information

Figure S1.1 relates to Figure 2.1, showing flow cytometry gating strategy

Figure S1.2 relates to Figure 2.1B, showing fluorescence microscopy of BCR_ABL-1 activators

Figure S1.3 relates to Figure 2.1C, showing fluorescence microscopy of all ZF activators

Figure S1.4 relates to Figure 2.1C, showing ZF activator orthogonality flow cytometry data

Figure S1.5 relates to Figure 2.1C, showing ZF activator orthogonality fluorescence microscopy

Figure S1.6 relates to Figure 2.2B, showing fluorescence microscopy of BCR_ABL-1 repressors

Figure S1.7 relates to Figure 2.2C, showing Fluorescence microscopy of all ZF repressors

Figure S1.8 relates to Figure 2.1C, showing ZF repressor orthogonality flow cytometry data

Figure S1.9 relates to Figure 2.1C, showing ZF repressor orthogonality fluorescence microscopy

Figure S1.10 relates to Figure 2.3A, showing OR gate flow cytometry data

Figure S1.11 relates to Figure 2.3A, showing OR gate fluorescence microscopy

Figure S1.12 relates to Figure 2.3B, showing NOR gate flow cytometry data

Figure S1.13 relates to Figure 2.3B, showing NOR gate fluorescence microscopy

Figure S1.14 relates to Figure 2.4D, showing ZF-TF split site fluorescence microscopy

Figure S1.15 relates to Figure 2.4, showing ZF-TF split site model structure

Figure S1.16 relates to Figure 2.5A, showing AND gate fluorescence microscopy

Figure S1.17 relates to Figure 2.5B, showing NAND gate fluorescence microscopy

Table S1.1 relates to Figures 2.1-2.5, showing transfected plasmids

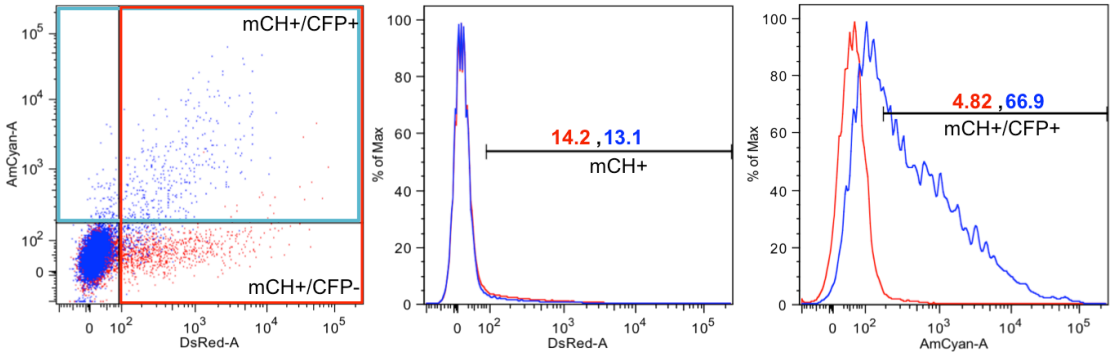
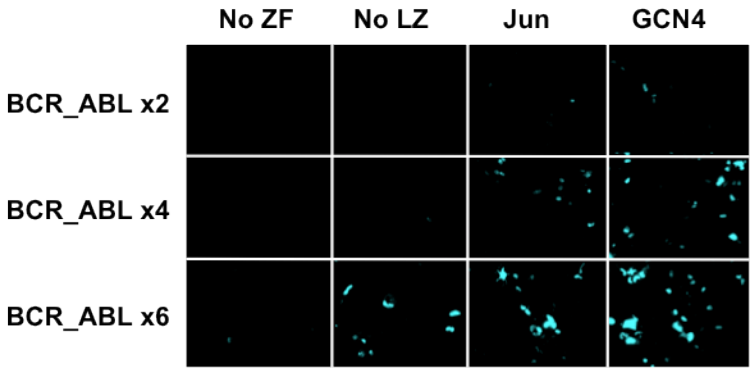


Figure S1.1 Example of raw data and overview of flow cytometry gating strategy

CFP vs. mCH scatter plot and histograms for BCR_ABL-1:GCN4 (blue dots and traces) and No ZF control (red dots and traces) co-transfection experiments with the 6x-BCR CFP reporter. The scatter plot shows live cells gated by mCH and CFP expression. The histograms show mCH+ and mCH+/CFP+ gating. Numbers represent the frequency of the cells that are mCH+ of total live cells (center) and cells that are CFP+ of total live mCH+ cells (right). mCH+ and CFP+ gates were determined by gating on a no DNA transfection control sample.

A



B

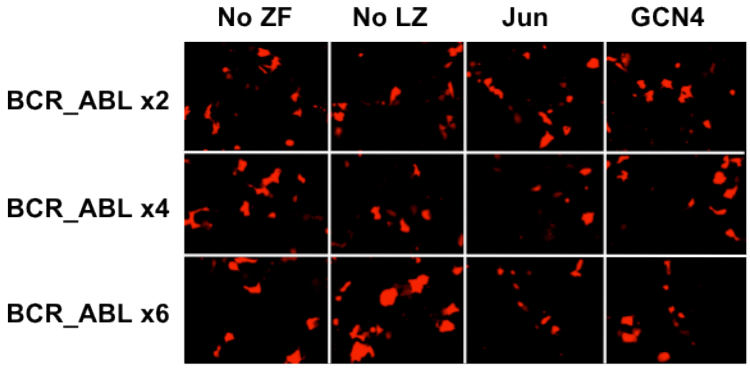
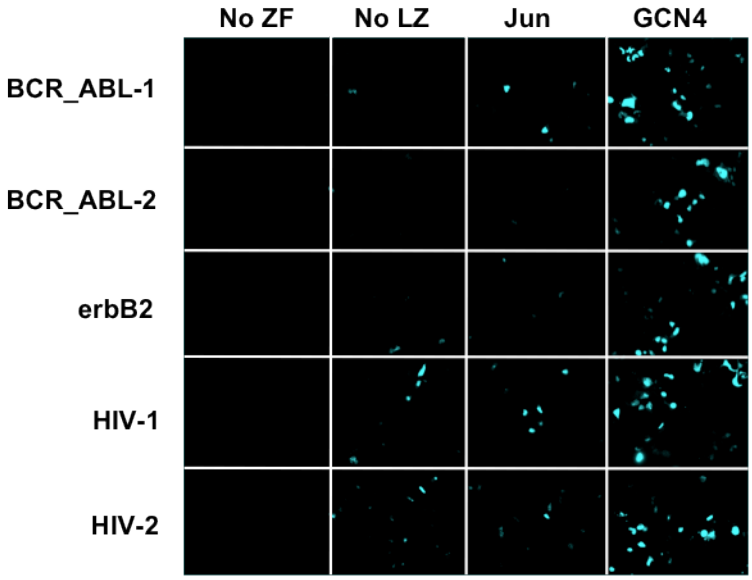


Figure S1.2 Fluorescence microscopy of BCR_ABL-1 activators

(A) Fluorescence microscopy images of BCR_ABL-1 activators driving cyan fluorescent protein (CFP) reporter expression. Each column shows U-2 OS cells transfected with a BCR_ABL-1 activator fused to either no leucine zipper (LZ), a c-Jun LZ (Jun), or a GCN4 LZ (GCN4). Each activator was co-transfected with a CFP reporter plasmid containing either 2, 4, or 6 copies of the 9bp BCR_ABL target site driving expression of 2x copies of CFP. (B) Fluorescence microscopy images of mCherry expression for the same cells as in A. As each activator construct was tagged with a t2A:mCherry fluorescent protein, these images mark transfected cells and confirm activator expression.

A



B

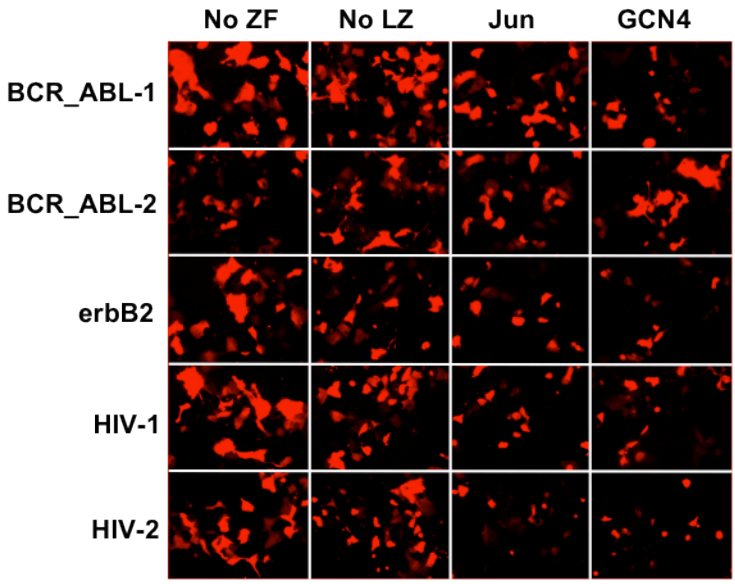


Figure S1.3 Fluorescence microscopy of all ZF-based activators

(A) Fluorescence microscopy images of all synthetic ZF-based activators driving CFP reporter expression. Cells were transfected with one of 5 ZF-based activators (BCR_ABL-1; BCR_ABL-2; erbB2; HIV-1; HIV-2) fused to either no leucine zipper (No LZ), a c-Jun LZ (Jun), or a GCN4 LZ (GCN4). Each activator was co-transfected with a CFP reporter plasmid containing 6 copies of the corresponding 9bp target site driving expression of 2x copies of CFP. (B) Fluorescence microscopy images of mCherry expression for the same cells as in A. confirming activator expression and marking transfected cells.

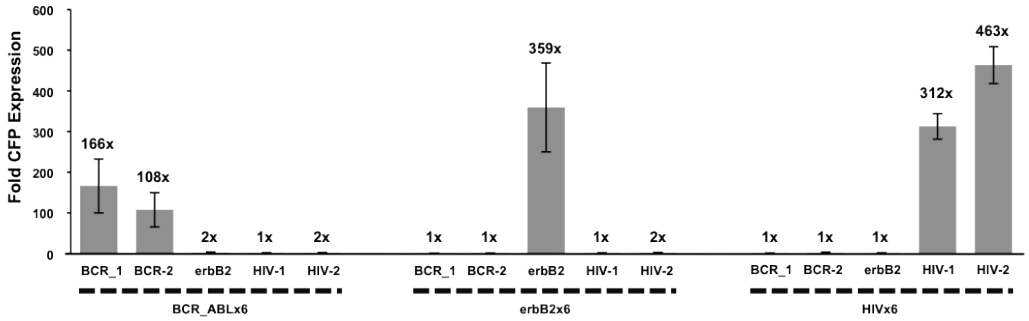
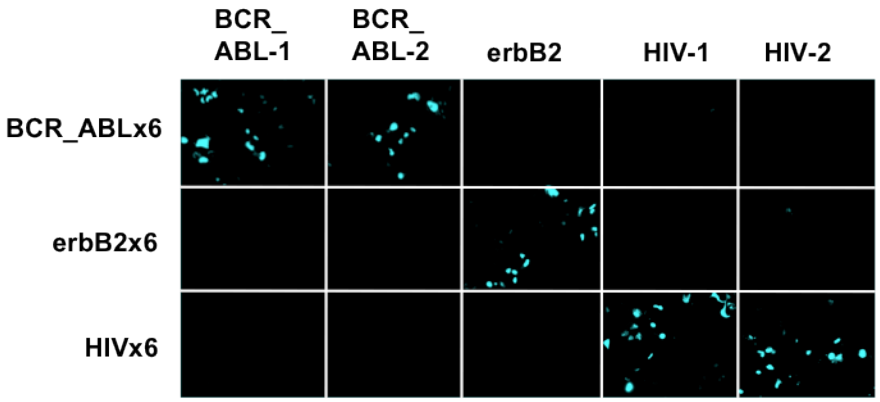


Figure S1.4 Flow cytometry data demonstrating ZF activator orthogonality

Flow cytometry data of cells transfected with one of 5 ZF-based activators (BCR_ABL-1; BCR_ABL-2; erbB2; HIV-1; HIV-2) fused to the GCN4 leucine zipper. Each activator was co-transfected with a CFP reporter plasmid containing 6 copies of the 9bp target site for either a BCR_ABL ZF, an erbB2 ZF, or an HIV ZF driving expression of 2x copies of CFP. Fold activations are calculated over total CFP values for co-transfections with each individual ZF reporter and an off-target control. All error bars indicate the standard deviation with n=3.

A



B

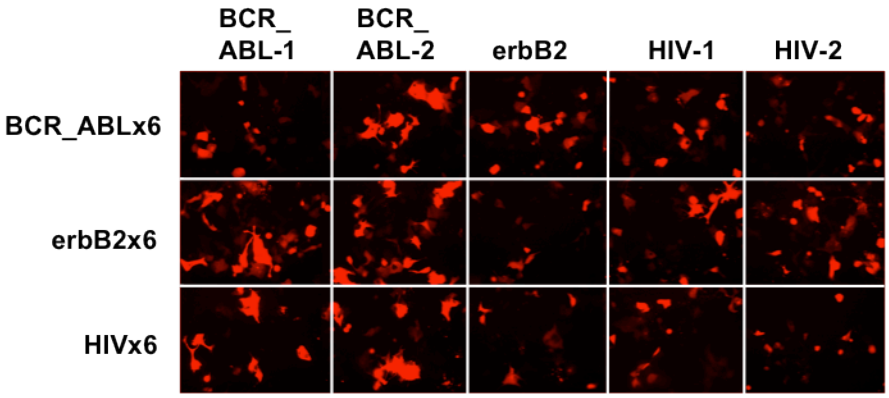
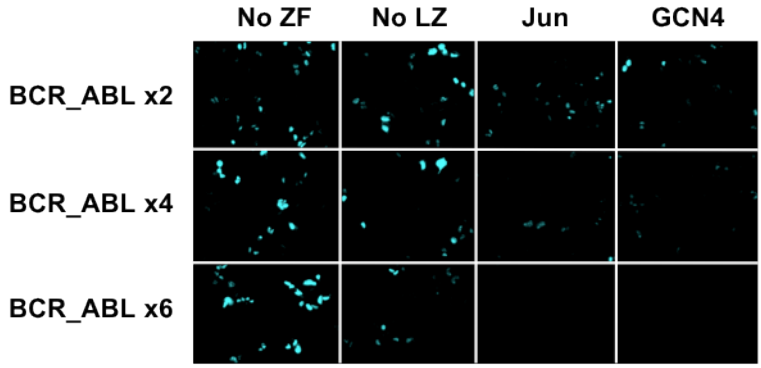


Figure S1.5 Fluorescence microscopy demonstrating the mutual orthogonality of ZF-based activators.

(A) Fluorescence microscopy images of synthetic ZF-based activators driving CFP reporter expression as described in Figure S1.3. (B) Fluorescence microscopy images of mCherry expression for the same cells as in A. confirming activator expression and marking transfected cells.

A



B

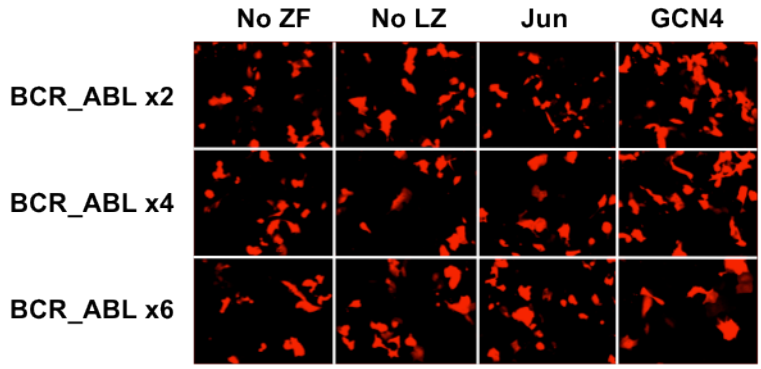


Figure S1.6 Fluorescence microscopy of BCR_ABL-1 repressors

(A) Fluorescence microscopy images of BCR_ABL-1-mediated CFP reporter repression. Cells were transfected with a BCR_ABL-1 repressor fused to either no leucine zipper (no LZ), a c-Jun LZ (Jun), or a GCN4 LZ (GCN4). Each repressor was co-transfected with a constitutively active CFP reporter plasmid containing either 2, 4, or 6 copies of the 9bp BCR_ABL target site inserted into the transcriptional start site of the CMV promoter driving expression of 2x copies of CFP. (B) Fluorescence microscopy images of mCherry expression for the same cells as in A. confirming repressor expression and marking transfected cells.

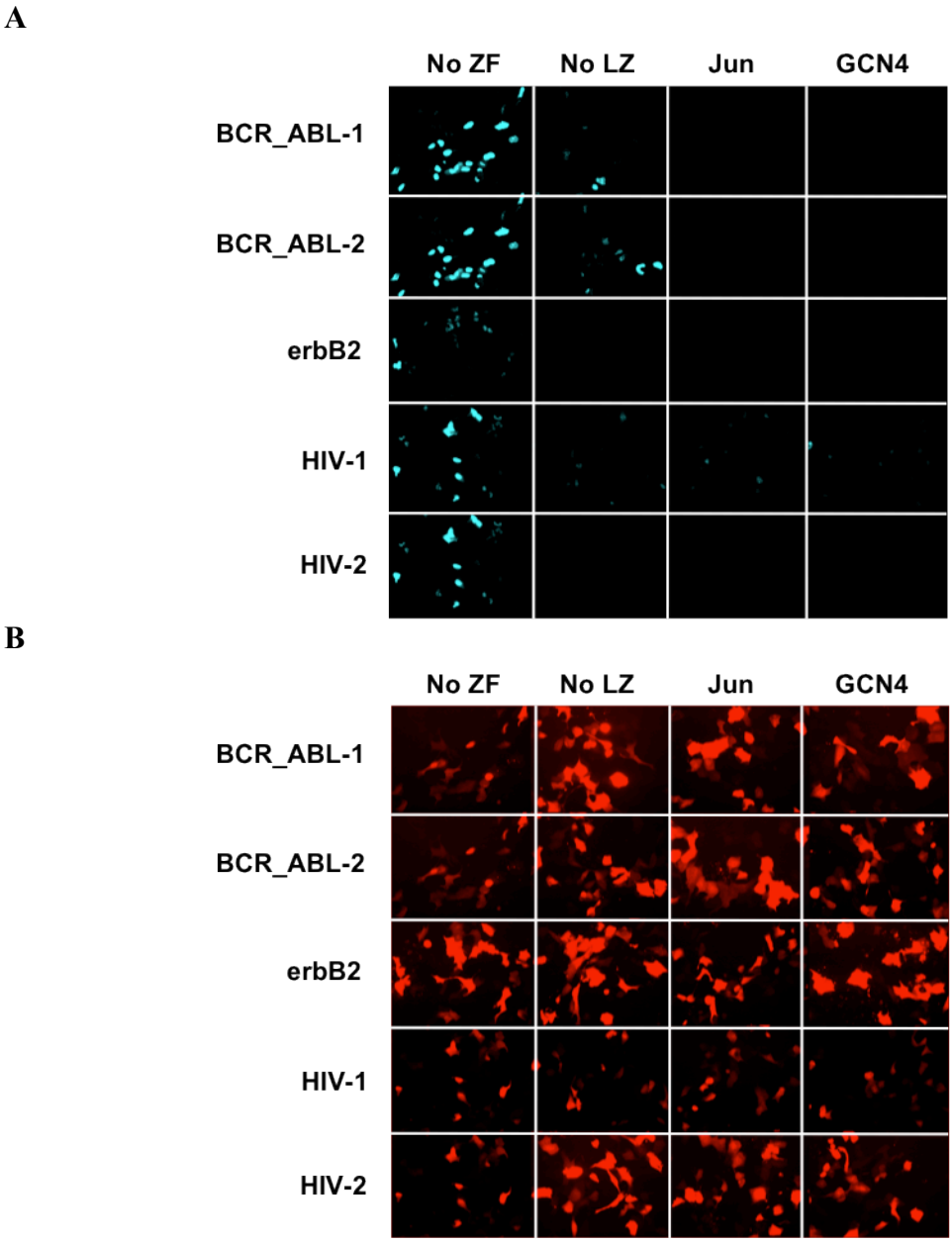


Figure S1.7 Fluorescence microscopy of all ZF-based repressors

(A) Fluorescence microscopy images of synthetic ZF repressors mediating CFP reporter repression. Cells were transfected with one of 5 ZF-based repressors (BCR_ABL-1; BCR_ABL-2; erbB2; HIV-1; HIV-2) fused to either no leucine zipper (no LZ), a c-Jun LZ (Jun), or a GCN4 LZ (GCN4). Each repressor was co-transfected with a CFP reporter plasmid containing 6 copies of the corresponding 9bp target site driving expression of 2x copies of CFP. The "no ZF" images shown for ZF pairs BCR_ABL-1 and BCR_ABL-2 and HIV-1 and HIV-2 are identical as each pair uses the same reporter construct. (B) Fluorescence microscopy images of mCherry expression for the same cells as in A. confirming repressor expression and marking transfected cells.

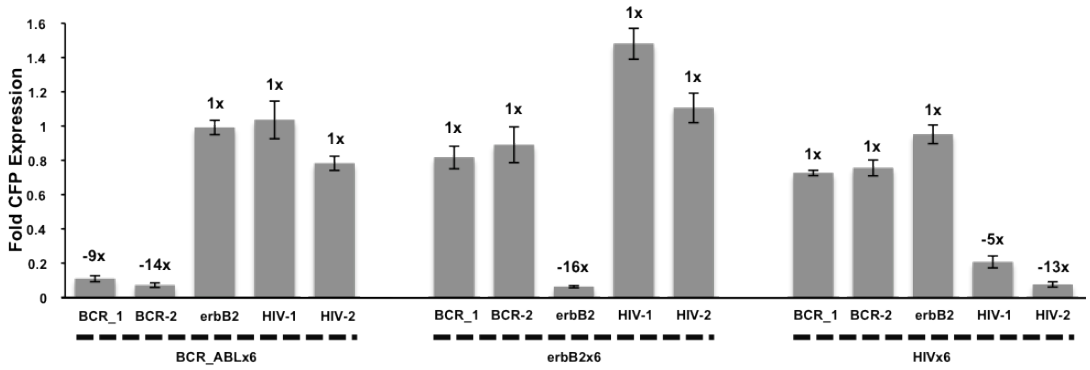
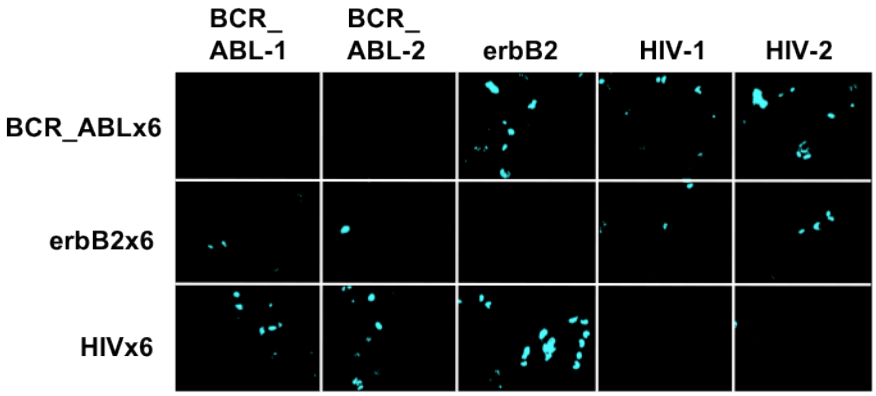


Figure S1.8 Flow cytometry data demonstrating ZF repressor orthogonality

Flow cytometry data of cells transfected with one of 5 ZF-based repressors (BCR_ABL-1; BCR_ABL-2; erbB2; HIV-1; HIV-2) fused to the GCN4 leucine zipper. Each repressor was co-transfected with a CFP reporter plasmid containing a repressible promoter with 6 copies of the 9bp target site for either a BCR_ABL ZF, an erbB2 ZF, or an HIV ZF driving expression of 2x copies of CFP. Fold activations are calculated over total CFP values for co-transfections with each individual ZF reporter and an off-target control. All error bars indicate the standard deviation with n=3.

A



B

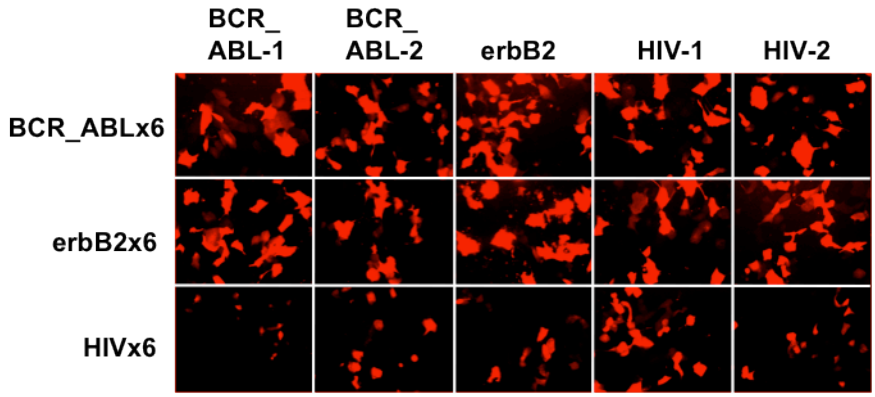


Figure S1.9 Fluorescence microscopy demonstrating the mutual orthogonality of ZF-based repressors

(A) Fluorescence microscopy images of synthetic ZF-based repressors driving CFP reporter expression as described in Figure S1.7. (B) Fluorescence microscopy images of mCherry expression for the same cells as in A. confirming repressor expression and marking transfected cells.

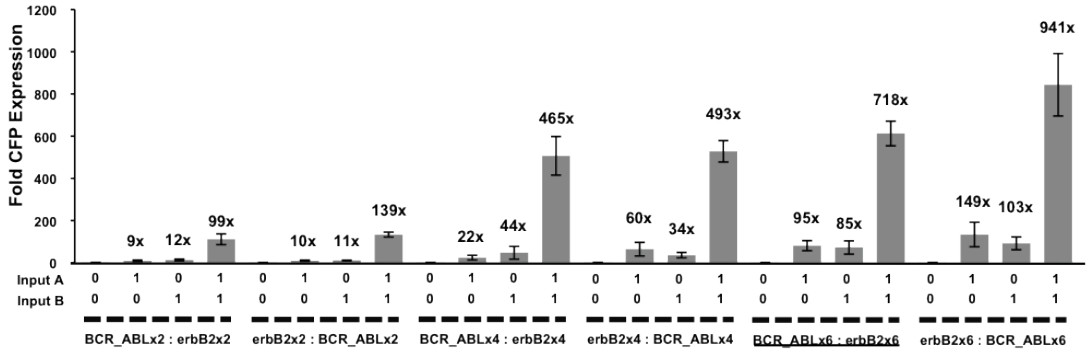
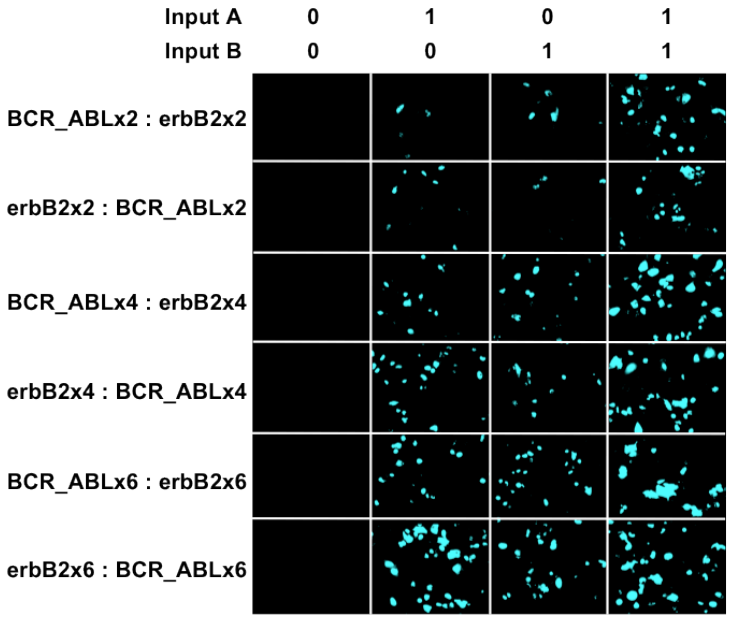


Figure S1.10 Characterization of ZF-based Boolean OR gates

OR gate reporter constructs were engineered with various numbers and combinations of ZF activator target sites. Each OR gate reporter construct was co-transfected with either a single corresponding ZF-activator alone, or with both activators present together. Input A represents the BCR_ABL-1:GCN4 activator, and input B represents the erbB2:Jun activator. Columns marked with 0 indicate that the corresponding input was not present in that experiment and was replaced bu the off-target ZF control (Zif268-t2a-mCherry). Columns marked with 1 indicate that the corresponding input was present in that experiment. All error bars indicate the standard deviation with n=3. The fold CFP expression was determined via flow cytometry analysis of mCherry positive cells and calculated as the ratio of total CFP for cells transfected with ZF-activator inputs to cells transfected with the off-target control.

A



B

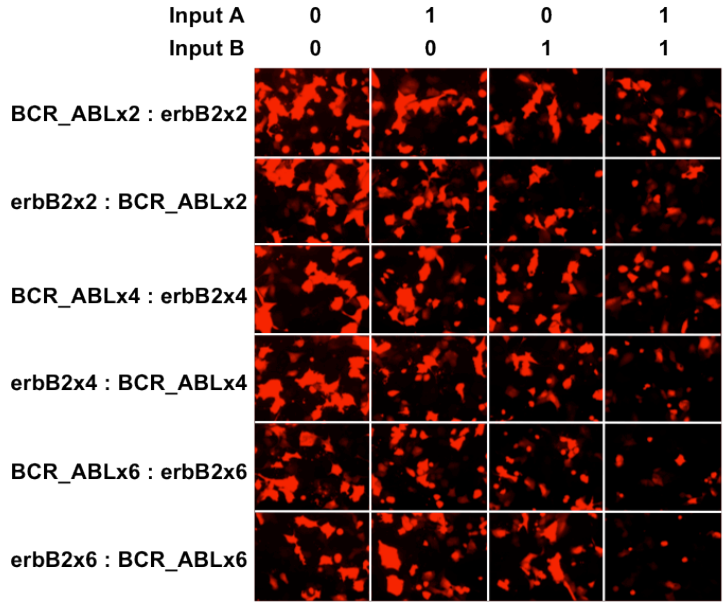


Figure S1.11 Fluorescence microscopy of ZF-based Boolean OR gates

(A) Fluorescence microscopy images of OR gate systems described in Figure S1.9. (B) Fluorescence microscopy images of mCherry expression for the same cells as in a. confirming input and off-target control expression and marking transfected cells.

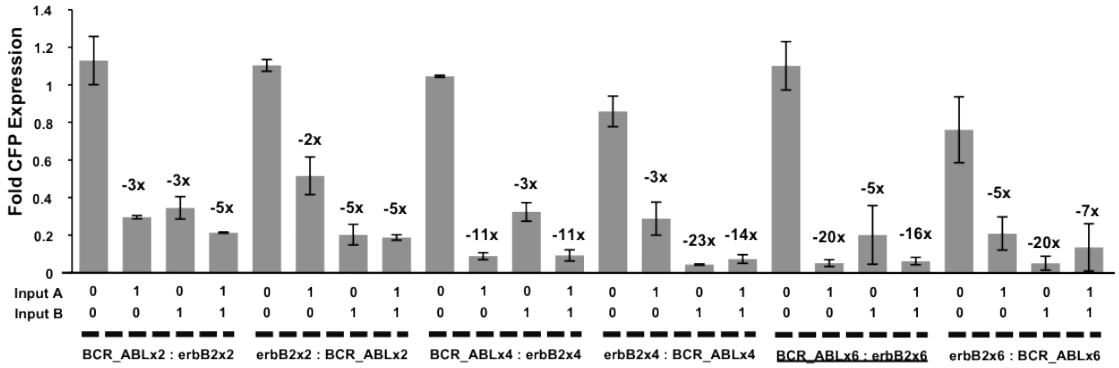
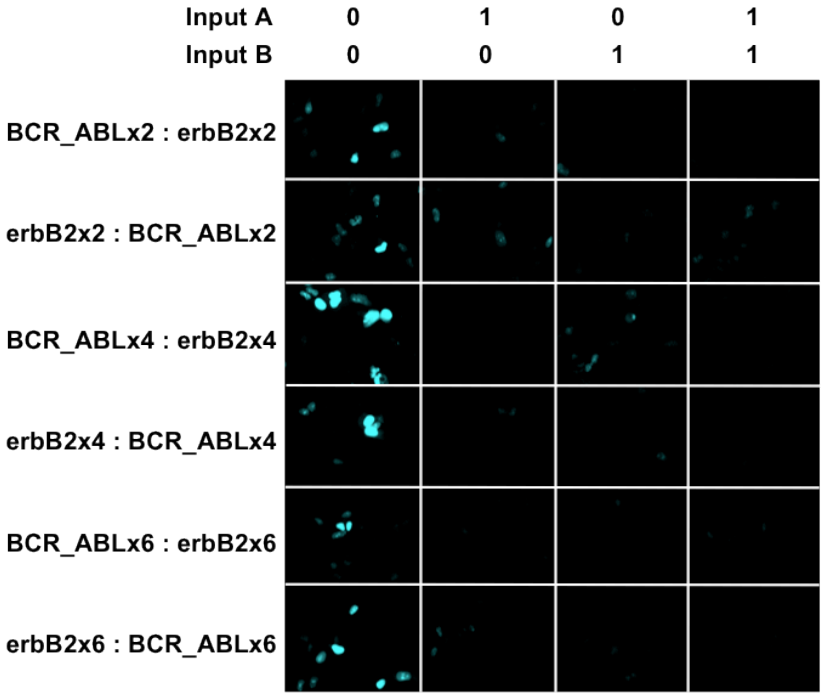


Figure S1.12 Characterization of ZF-based Boolean NOR gates

NOR gate reporter constructs were engineered with various numbers and combinations of ZF repressor target sites inserted into the transcriptional start site for 2x CFP. Each NOR gate reporter construct was co-transfected with either a single corresponding ZF-repressor alone, or with both repressors together. Input A represents the BCR_ABL-1:GCN4 repressor, and input B represents the erbB2:Jun repressor. Columns marked with 0 indicate that the corresponding input was not present in that experiment, and was replaced with the off-target ZF control (Zif268-t2a-mCherry). Columns marked with 1 indicate that the corresponding input was present in that experiment. All error bars indicate the standard deviation with n=3. The fold CFP expression was determined via flow cytometry analysis of mCherry positive cells and calculated as the ratio of total CFP for cells transfected with ZF-repressor inputs to cells transfected with a non-DNA binding control.

A



B

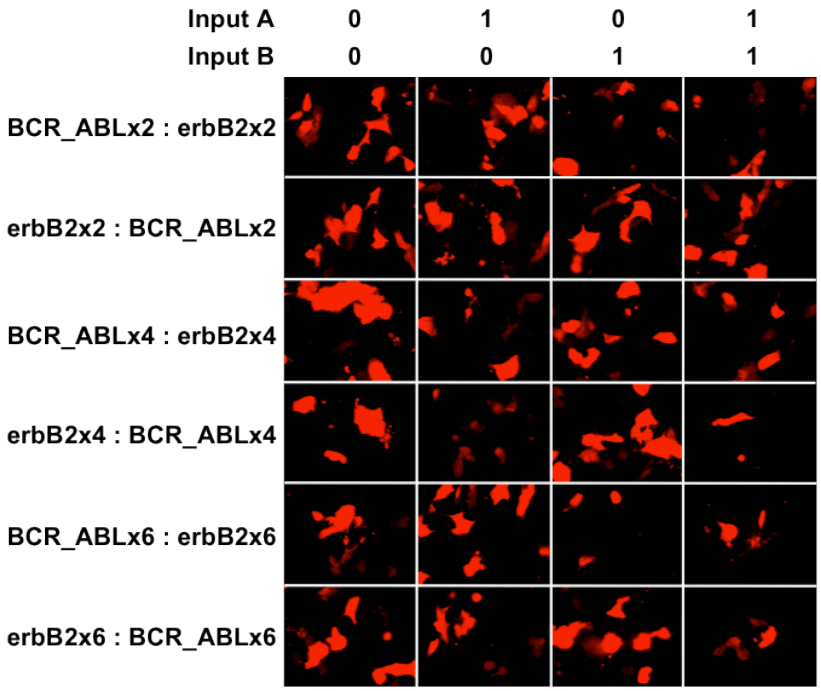


Figure S1.13 Fluorescence microscopy of ZF-based Boolean NOR gates

(A) Fluorescence microscopy images of synthetic NOR gate systems described in Figure S1.11. (B) Fluorescence microscopy images of mCherry expression for the same cells as in A. confirming input and off-target control expression and marking transfected cells.

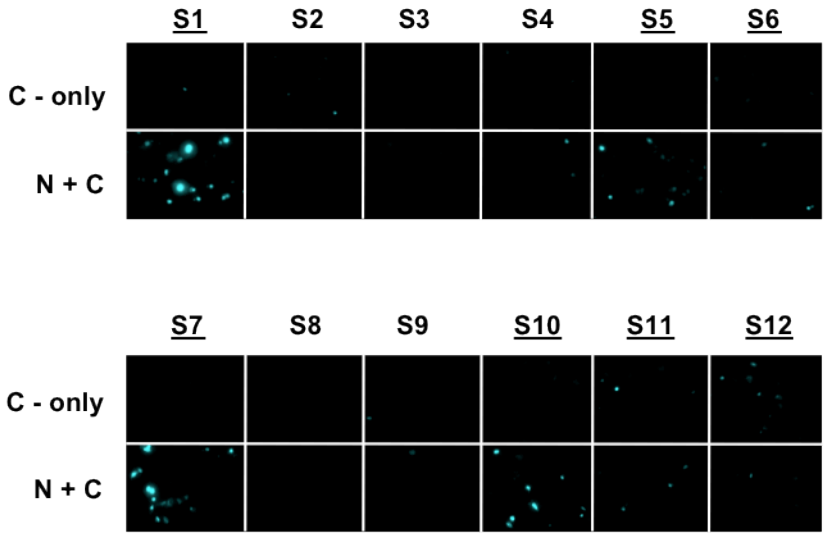


Figure S1.14 Fluorescence microscopy of ZF-TF split sites

Fluorescence microscopy images of the split-intein ZF activator split site assay described in Figure S1.13. Images are shown for the Split pairs and the C-fragment alone. Split sites that lead to a >3 fold activation over the C-fragment alone are underlined.

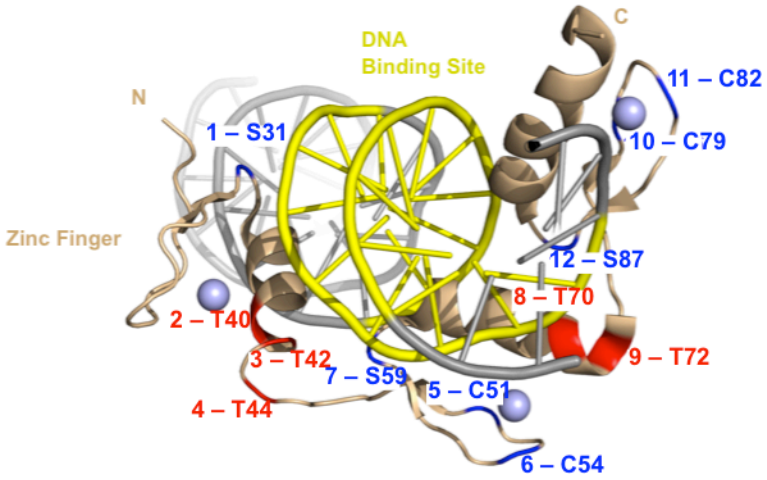
A

```
# Identity:      49/87 (56.3%)
# Similarity:    61/87 (70.1%)
# Gaps:           2/87 ( 2.3%)


Zif268          103 RPYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCRICMRNFSRSDHLTT     153
                    ||:.|.:  |.|.||.|..|.||.|.|||:|||||||||||||:..:|..
BCR_ABL-1        19 RPFQCRI--CMRNFSDSPTLRRHTRTHTGEKPFQCRICMRNFSQGANLRR     67

Zif268          153 HIRTHTGEKPFACDICGRKFARSDERKRHTKIHLRQ   186
                    |:||||||||.|.||.|.|::::..:||.|.|..:
BCR_ABL-1        68 HLRTHTGEKPFQCRICMRNFSQANTLQRHLKTHTGE   102
```

B

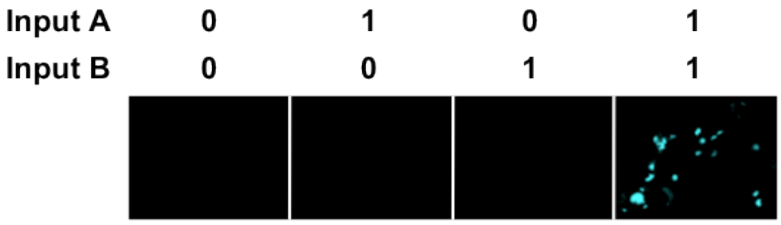



```
Split Sites                 1       2 3 4     5  6   7        8 9      10 11   12
BCR_ABL1 Seq    …RPFQCRICMRNFSDSPTLRRHTRTHTGEKPFQCRICMRNFSQGANLRRHLRTHTGEKPFQCRICMRNFSQANTLQRHLKTHTGE…
(aa 19-102)
```

Figure S1.15 Zinc finger structure overlaid with split sites

(A) Amino Acid sequence alignment of BCR_ABL-1 (amino acid residues 19-102) with the sequence of Zif268 (amino acid residues 103-186) from PDB 1A1L and alignment statistics generated using the 'needle' function from the EBI Tools Package (Laberga et al. 2007). (B) The ZF split sites assayed for splicing are labeled in the primary sequence of BCR_ABL1 and highlighted on the crystal structure of Zif268 PDB ID 1A1L (Elrod-Erickson et al. 2003). Residues in blue are split sites with splicing efficiencies of >3 fold over the C-fragment alone, and residues in red are split sites with no significant splicing activity. Effective split sites are all within protein loop regions.

A



B

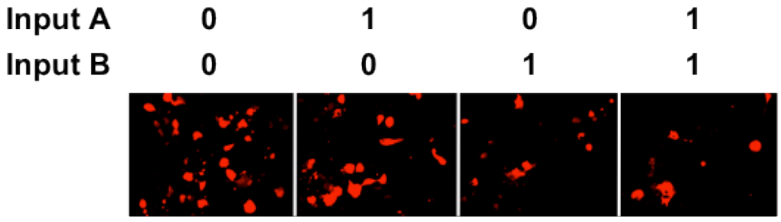
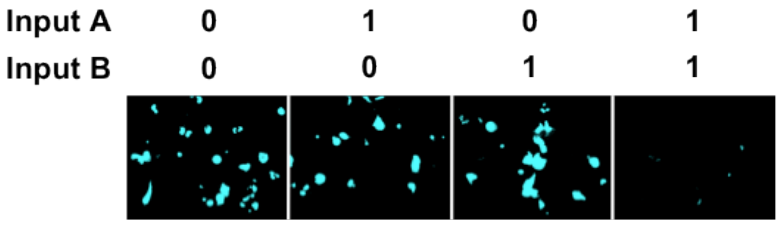


Figure S1.16 Fluorescence microscopy of ZF-based Boolean AND gates

(A) Fluorescence microscopy images of AND gate induced CFP reporter expression from BCR_ABL activator reporter constructs. Cells were transfected with the 6x BCR_ABL activator reporter along with either input A alone, input B alone, or both inputs present together. Here, Input A is the modified N-terminal BCR_ABL-1:GCN4 :intein* activator fragment, and input B is the C-terminal BCR_ABL-1:GCN4:intein activator fragment. Columns marked with 0 indicate that the corresponding input was not present in that experiment, whereas columns marked with 1 indicate that the corresponding input was present in that experiment. (B) Fluorescence microscopy images of mCherry expression for the same cells as in A. confirming input and off-target control expression and marking transfected cells.

A



B

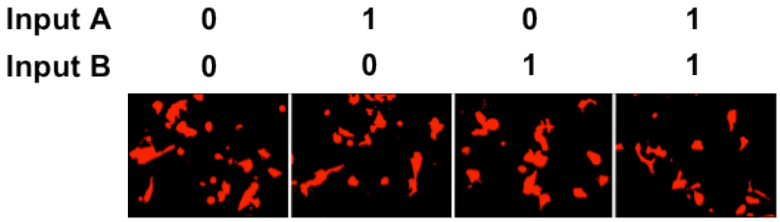


Figure S1.17 Fluorescence microscopy of ZF-based Boolean NAND gates

(A) Fluorescence microscopy images of NAND gate induced CFP reporter repression from BCR_ABL repressor reporter constructs. U-2 OS cells were transfected with the 6x BCR_ABL repressor reporter along with either input A alone, input B alone, or both inputs present in tandem. Here, Input A is the modified N-terminal BCR_ABL-1:GCN4:intein* repressor fragment, and input B is the C-terminal BCR_ABL-1:GCN4:intein repressor fragment. Columns marked with 0 indicate that the corresponding input was not present in that experiment, whereas columns marked with 1 indicate that the corresponding input was present in that experiment. (B) Fluorescence microscopy images of mCherry expression for the same cells as in A. confirming input and off-target control expression and marking transfected cells.

Table S1.1 DNA plasmids co-transfected for each ZF experiment

Plasmids and plasmid amounts transfected for each activator, repressor, and logic gate experiment. These amounts were chosen based on preliminary titration experiments. Note that transcription factor and reporter amounts differ between the activator and repressor experiments. The repressor systems have a high ZF-TF:reporter plasmid ratio as in the transient assays the repressible promoter has a 'head-start' to produce CFP protein before sufficient repressor is made to repress CFP expression.

Activators	**ZF Construct** *Activator Expression Construct*	**Off Target** *intC-BCR1-GCN4-C1*	**Reporter** *Activator Reporter*	**Empty Vector** *pCDNA5-ins*
No ZF	--	10ng	990ng	0ng
ZF	10ng	--	990ng	0ng

Repressors	**ZF Construct** *Repressor Expression Construct*	**Off Target** *intC-BCR1-GCN4-C1*	**Reporter** *Repressor Reporter*	**Empty Vector** *pCDNA5-ins*
No ZF	--	100ng	10ng	890ng
ZF	100ng	--	10ng	890ng

OR		**Input A**	**Input B**	**Off Target**	**Reporter**	**Empty Vector**
Input A	Input B	*Activator: BCR_ABL-1-GCN4*	*Activator: erbB2-Jun*	*Off-target ZF*	*OR gate Reporter*	*pCDNA5-ins*
0	0	--	--	20 ng	980 ng	0 ng
1	0	10 ng	--	10 ng	980 ng	0 ng
0	1	--	10 ng	10 ng	980 ng	0 ng
1	1	10 ng	10 ng	--	980 ng	0 ng

NOR		**Input A**	**Input B**	**Off Target**	**Reporter**	**Empty Vector**
Input A	Input B	*Repressor: BCR_ABL-1-GCN4*	*Repressor: erbB2-Jun*	*Off-target ZF*	*NOR gate Reporter*	*pCDNA5-ins*
0	0	--	--	200 ng	10 ng	790 ng
1	0	100 ng	--	100 ng	10 ng	790 ng
0	1	--	100 ng	100 ng	10 ng	790 ng
1	1	100 ng	100 ng	--	10 ng	790 ng

Table S1.1 (Continued).

AND		**Input A**	**Input B**	**Off Target**	**Reporter**	**Empty Vector**
Input A	Input B	*BCR_ABL1-N1-IntN**	*Activator: intC-BCR_ABL1-C1-GCN4*	*Off-target ZF*	*Activator: 6x-BCR_ABL*	*pCDNA5-ins*
0	0	--	--	100 ng	900 ng	0 ng
1	0	50 ng	--	50 ng	900 ng	0 ng
0	1	--	50 ng	50 ng	900 ng	0 ng
1	1	50 ng	50 ng	--	900 ng	0 ng

NAND		**Input A**	**Input B**	**Off Target**	**Reporter**	**Empty Vector**
Input A	Input B	*BCR_ABL1-N1-IntN**	*Repressor: intC-BCR_ABL1-GCN4*	*Off-target ZF*	*Repressor: 6x-BCR_ABL*	*pCDNA5-ins*
0	0	--	--	200 ng	10ng	790 ng
1	0	100 ng	--	100 ng	10ng	790 ng
0	1	--	100 ng	100 ng	10ng	790 ng
1	1	100 ng	100 ng	--	10ng	790 ng

ZF Splits	**N-Fragment**	**C-Fragment**	**Reporter**	**Empty Vector**
	BCR1-N#-intN	*Activator: intC-BCR1-C#-Jun*	*Activator: 6x-BCR_ABL*	*pCDNA5-ins*
C-only	--	150	700ng	150ng
N+C	150ng	150ng	700ng	150ng

SUPPLEMENTAL REFERENCES

Labarga, A., Valentin, F., Anderson, M. and Lopez, R. (2007) Web services at the European bioinformatics institute. Nucleic Acids Res, 35, W6-11.

Elrod-Erickson, M., Benson, T.E. and Pabo, C.O. (1998) High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. Structure, 6, 451-464.

Appendix II

Supplemental Information: Engineering synthetic TAL effectors with orthogonal target sites

Abhishek Garg[1], Jason J Lohmueller[1], Pamela A Silver[1,2], and Thomas Z Armel[1]

[1]*Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA*

[2]*Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts 02115, USA*

Reproduced from Garg A, Lohmueller JJ, Silver PA, Armel TZ. (2012). Engineering synthetic TAL effectors with orthogonal target sites..*Nucleic Acids Res.* Aug;40(15):7584-95.

Abhishek Garg contributed data for Supplemental Information in collaboration with Jason J. Lohmueller and contributions from Thomas Z. Armel.

Inventory of Supplemental Information

Figure S2.1 relates to Chapter III, showing reverse-triangle inequality heuristic

Figure S2.2 relates to Chapter III, showing TALE cloning strategy

Figure S2.3 relates to Figure 3.5, showing TALE expression

Figure S2.4 relates to Chapter III, showing the distribution of 18bp TALE binding sites in the human genome

Figure S2.5 relates to Chapter III, showing the distribution of 20bp TALE binding sites in the human genome

Figure S2.6 relates to Chapter III, showing a ROBDD representation

Figure S2.7 relates to Chapter III, showing an ADD of the hamming distance

Table S2.1 relates to Figure 3.5, showing fold induction of CFP reporter by synthetic TALEs

Table S2.2 relates to Chapter III, showing optimal target sequence of synthetic TALEs and their closest endogenous target sequence in 2000 bp promoter regions

Table S2.3 relates to Chapter III, showing the position of mismatches between the optimal target sequences of the synthetic TALEs

Table S2.4 relates to Table 3.1, showing subparts used to assemble each of the synthetic TALEs

Table S2.5 relates to Figure 3.7, showing target DNA sequences of TALEs in the endogenous promoter regions

Table S2.6 relates to Figure 3.7, showing reverse and forward strand primer sequences used in qPCR experiments

Table S2.7 relates to Figure 3.8, showing constructs used in TALE repressor-shRNA co-expression experiments

Supplementary Methods relates to Chapter III.

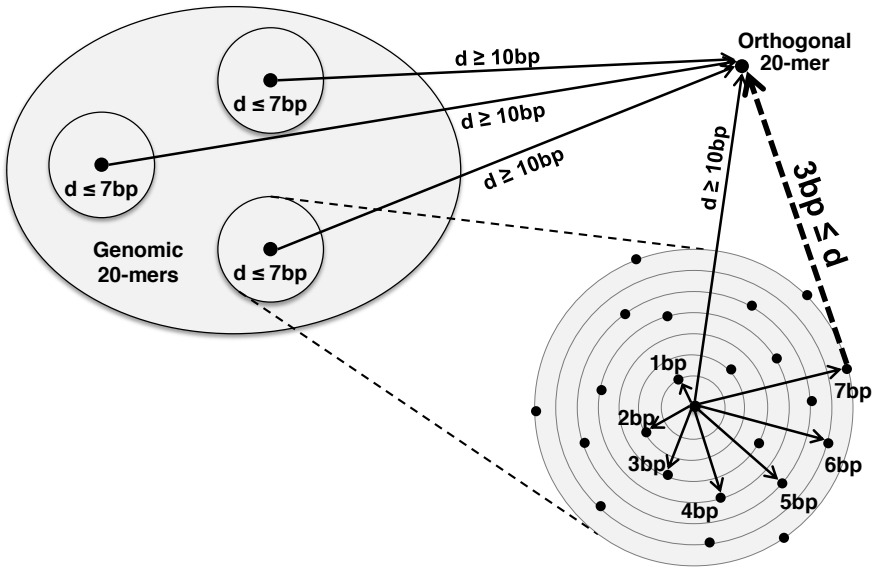
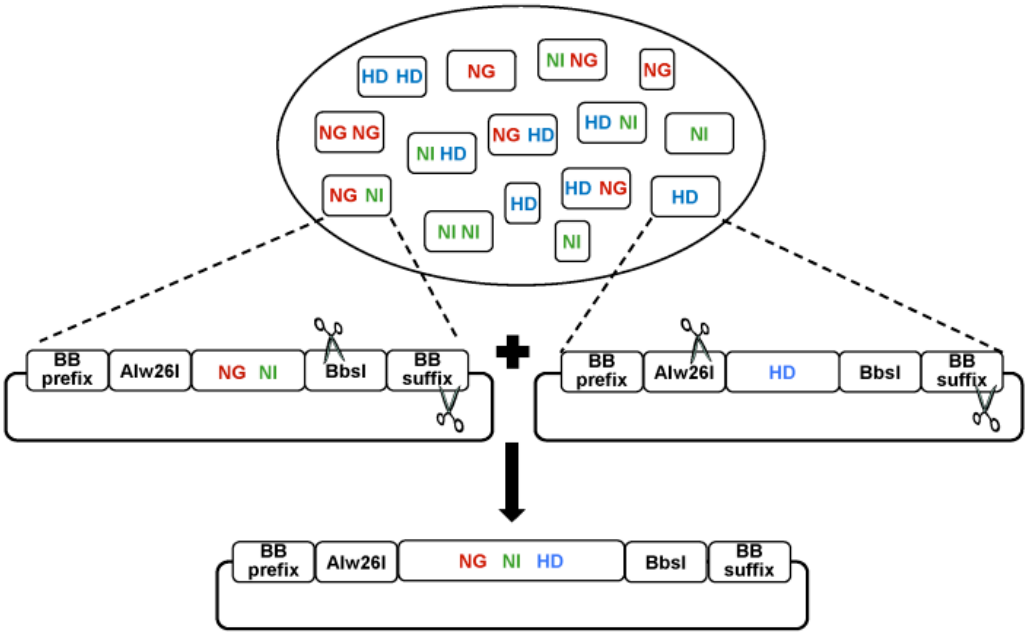


Figure S2.1 Schematic representation of the reverse-triangle inequality heuristic used to generate orthogonal 20-mers

The complete set of genomic 20-mers is divided into smaller subsets (circles) and represented by a single sequence (central point). The zoomed-in subset shows individual sequences (points) are all 7bp or closer to the representative sequence (central point). Orthogonal sequences that are at a minimum 10 bp from the representative sequence (central point) are at least 3bp away from every sequence in the subset.

(A)



(B)

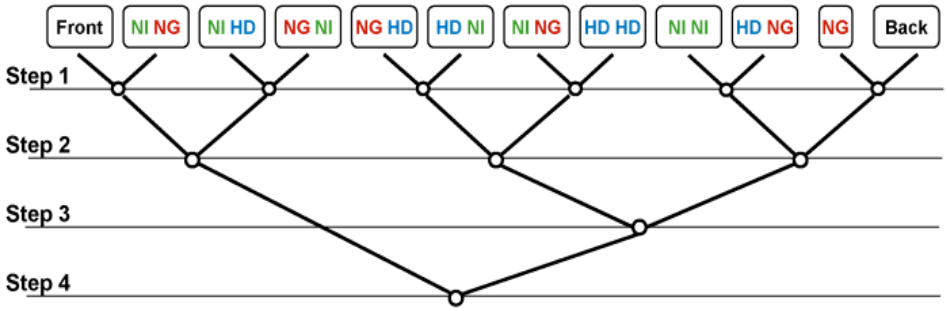


Figure S2.2 TALE cloning strategy

(A) TALE proteins were assembled from a pool of sub-parts that include repeat regions for RVDs and half RVDs targeting nucleotides A, C and T, and RVD pairs targeting all combinations of A, C, and T. TALE subparts were assembled by digesting the plasmid containing the 5'- RVD with the restriction enzymes BbsI and PstI and the plasmid containing 3'- RVD with the restriction enzymes Alw26I and PstI. The two DNA fragments were then ligated to generate a construct containing both RVD domains while also reconstituting the flanking enzyme sites. The BbsI and Alw26I sticky ends combine to generate the first four bases of the second RVD domain, thus leaving no nucleotide scar between ligated RVD domains. (B) Hierarchical assembly of a representative synthetic TALE. The small circles associated with each step represent the ligation of two parts. Using this method, a synthetic TALE engineered to recognize a 20 bp target sequence can be efficiently assembled in 4 cloning steps.

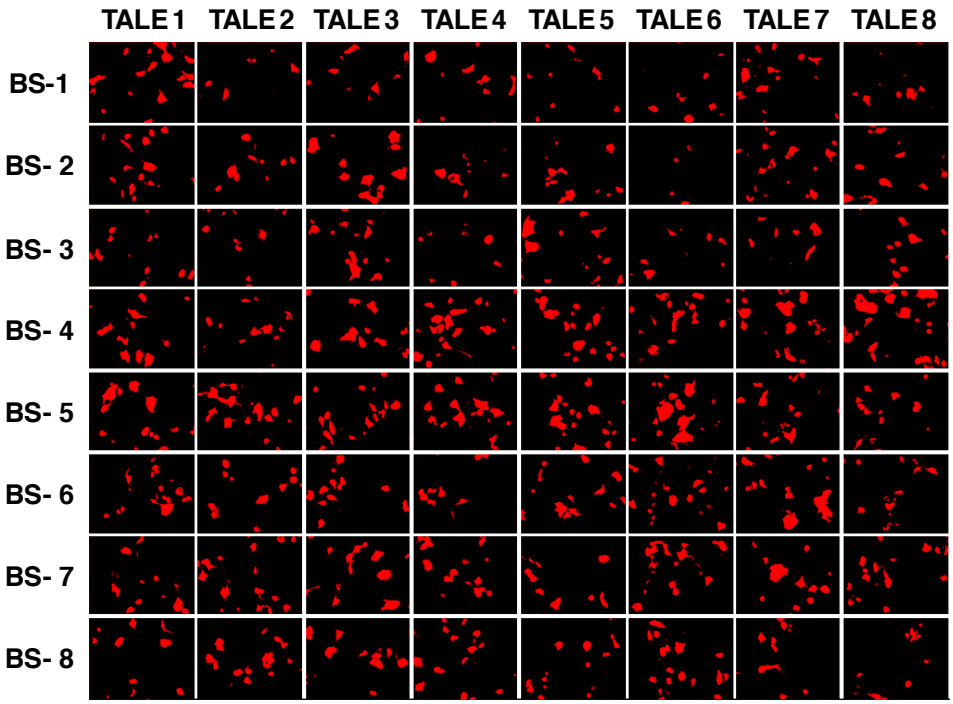


Figure S2.3 Fluorescent microscopy images of mCherry expression as a marker for TALE expression

Each column of the 8x8 matrix represents U-2OS cells co-transfected with a synthetic TALE and reporter constructs for each 20-mer binding site (BS). All TALE expression constructs were tagged with autocatalytically cleaved t2A:mCherry as a marker for protein expression. The presence of mCherry fluorescence in each well indicates that our TALE constructs are efficiently expressed in all samples.

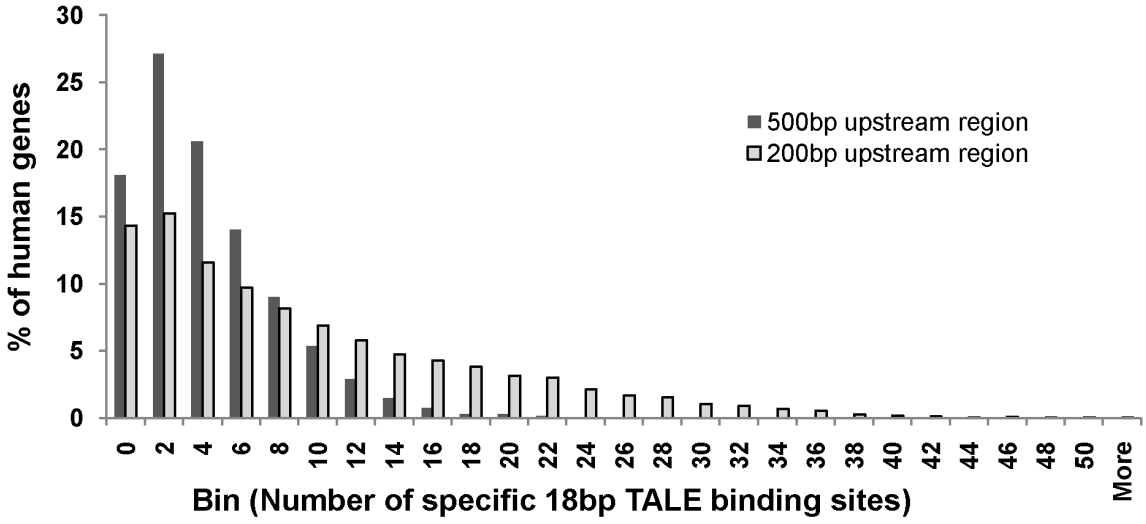


Figure S2.4 The distribution of promoter specific 18 bp TALE binding sites in the human genome

The bins on the X-axis represent the number of sites in a single 200 bp upstream region or 500 bp upstream region for which a TALE targeting that specific promoter region can be designed. The Y-axis represents the corresponding percentage of human genes that fall into each bin.

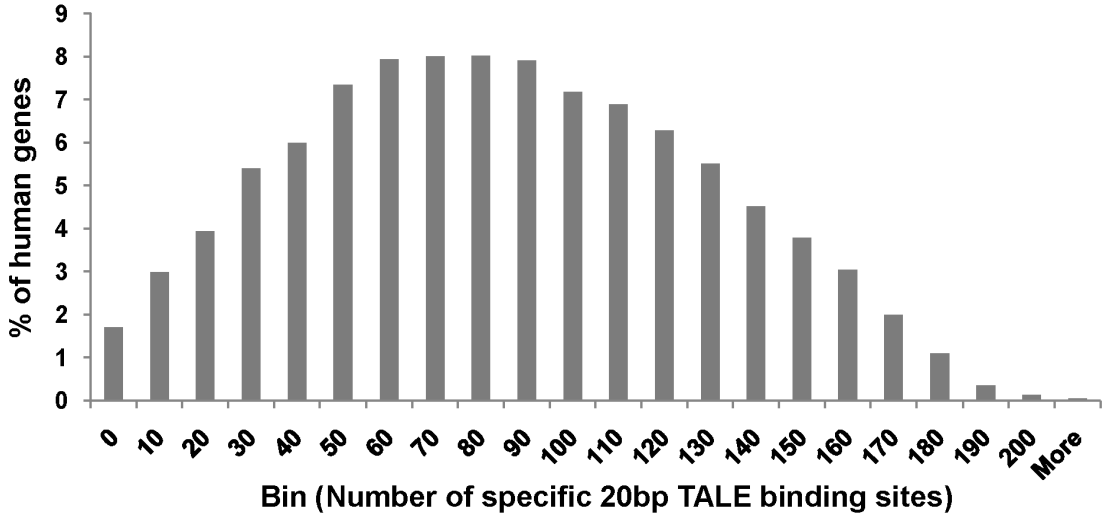


Figure S2.5 The distribution of promoter specific 20 bp TALE binding site in the human genome

The bins on the X-axis represent the number of sites in a single 500 bp upstream region for which a TALE targeting that specific promoter region can be designed. The Y-axis represents the corresponding percentage of human genes that fall into each bin.

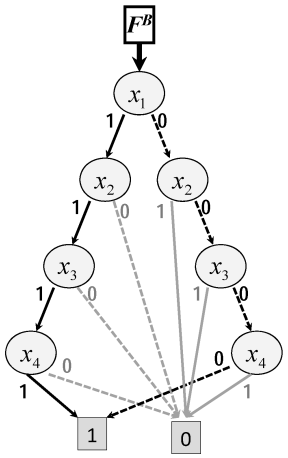


Figure S2.6 A ROBDD representation of the Boolean function
$F^B = [(x]_1 \wedge x_2 \wedge x_3 \wedge x_4) \vee [(\bar{x}]_1 \wedge \bar{x}_2 \wedge \bar{x}_3 \wedge \bar{x}_4)$.

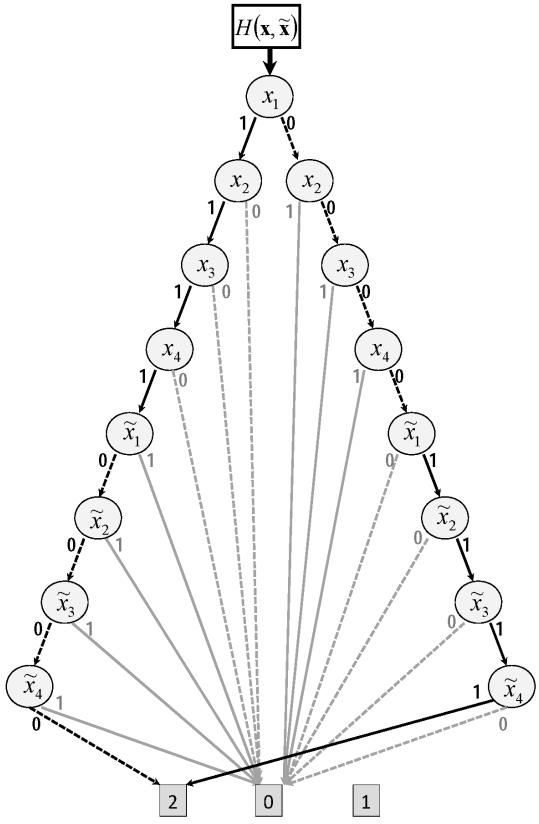


Figure S2.7 An ADD of the hamming distance between a set of 2-mers {TT,AA} in a Boolean representation using Equation S2

Table S2.1 Fold induction of CFP reporter by synthetic TALEs

Each column of the 8x8 matrix represents U2-OS cells co-transfected with a synthetic TALE and reporter constructs for each 20-mer binding site (BS). The level of CFP expression in each sample was measured by flow cytometry and the fold induction was calculated for each binding site relative to transfection with an off-target reporter. The asterisk indicates a sample for which only a single repeat is present due to machine error.

	TALE 1	**TALE 2**	**TALE 3**	**TALE 4**	**TALE 5**	**TALE 6**	**TALE 7**	**TALE 8**
BS-1	**20.89**	1.23	1.57	1.71	1.56	0.85	1.29	1.22
BS-2	0.93	**9.82**	0.98	0.89	0.98	1.01	0.81	0.79
BS-3	1.01	1.36	**65.67**	1.17	1.17	1.27	1.09	0.99
BS-4	0.86	0.82	0.74*	**10.2**	1.02	0.9	0.79	0.78
BS-5	0.81	1.15	1.09	1.58	**58.4**	0.77	1.21	1.04
BS-6	0.77	0.88	0.79	1.06	1	**101.68**	1.01	0.83
BS-7	1.14	0.99	0.81	1.34	1.25	0.98	**74.15**	1.1
BS-8	0.96	0.98	1.21	1.12	0.92	1.1	0.9	**85.71**

Table S2.2 Optimal target sequence of synthetic TALEs used in this paper and their closest endogenous target sequence in 2000 bp promoter regions

The optimal target sequence and the endogenous sequence that has the minimum hamming distance to the optimal target sequence are listed along with the number of mismatches between the two sequences. The position of the mismatches between the two sequences are listed numerically from 5'→3'. The first base pair position corresponding to 'T' is considered as 0.

TALE ID	Target Sequence (5'→3')	Endogenous Sequence (5'→3')	Number of Mismatch	Positions of Mismatch		
TALE-1	TCAATACTTACAAACTCC	TCAATGTTTATAAACTCC	3	5	6	10
TALE-2	TCCACCAAATTCAACACT	TCAAGCAACTTCAACACT	3	2	4	8
TALE-3	TCATCTACAACACTACTA	TCATCTACAACCCTGCAA	3	11	14	16
TALE-4	TCCCAATACACTATAACA	TCACAACACACTATAAGA	3	2	6	16
TALE-5	TAACTTACCTTCTCAACA	TAACGTACCTTGTCACCA	3	4	11	15
TALE-6	TATCCTCTTACAATATCC	TATCCTCTTATTACATCCC	3	10	11	13
TALE-7	TACTTACCCTAACCCAAT	TACTTACCCTATACAAAT	3	11	12	14
TALE-8	TATACTATCCAATCCAAC	TACACCATCCAATCCATC	3	2	5	16

Table S2.3 The position of mismatches between the optimal target sequence of the synthetic TALEs and their endogenous target sequence

The longest stretch of sequence from the 5'-end without any mismatches found in a 2000 bp upstream promoter region in the human genome is listed. The position the mismatches are ordered numerically from 5'→3'. The first base pair position corresponding to 'T' is considered as 0.

TALE ID	Optimal Target Sequence (5'→3')	Endogenous Sequence (5'→3')	Positions of Mismatch						
TALE-	TCAATACTTACAAACTCCTT	TCAATACTTACAGCCCACAC	12	13	15	16	18	19	
TALE-	TCCACCAAATTCAACACTTT	TCCACCAAATTCACAAAACC	13	14	16	17	18	19	
TALE-	TCATCTACAACACTACTATT	TCATCTACAACACAGAGTTG	13	14	15	16	17	19	
TALE-	TCCCAATACACTATAACACA	TCCCAATACACTCTGCCTCC	12	14	15	17	19		
TALE-	TAACTTACCTTCTCAACACA	TAACTTACCTTCTATGCTAA	13	14	15	17	18		
TALE-	TATCCTCTTACAATATCCCA	TATCCTCTTACATTAAAGCA	12	15	16	17			
TALE-	TACTTACCCTAACCCAATTT	TACTTACCCTAAGCTCTCCC	12	14	15	16	17	18	19
TALE-	TATACTATCCAATCCAACTT	TATACTATCCAAAGGCCACT	12	13	14	15	16	17	18

Table S2.4 Subparts used to assemble each of the synthetic TALEs used in this study

TALE ID	**Subpart composition (5'→3')**												
TALE 1	HD-L	NI-NI	NG-NI	HD-NG	NG-NI	HD-NI	NI-NI	HD-NG	HD-HD	NG-L	NG-S		
TALE 2	HD-L	HD-NI	HD-HD	NI-NI	NI-L	NG-NG	HD-NI	NI-L	HD-NI	HD-NG	NG-L	NG-S	
TALE 3	HD-L	NI-L	NG-HD	NG-NI	HD-NI	NI-L	HD-NI	HD-NG	NI-L	HD-NG	NI-L	NG-L	NG-S
TALE 4	HD-L	HD-L	HD-NI	NI-L	NG-NI	HD-NI	HD-NG	NI-L	NG-NI	NI-L	HD-NI	HD-L	NI-S
TALE 5	NI-L	NI-L	HD-NG	NG-NI	HD-L	HD-NG	NG-HD	NG-HD	NI-L	NI-L	HD-NI	HD-L	NI-S
TALE 6	NI-L	NG-HD	HD-NG	HD-NG	NG-NI	HD-NI	NI-L	NG-NI	NG-HD	HD-HD	NI-S		
TALE 7	NI-L	HD-NG	NG-NI	HD-HD	HD-NG	NI-NI	HD-HD	HD-NI	NI-L	NG-NG	NG-S		
TALE 8	NI-L	NG-NI	HD-NG	NI-L	NG-HD	HD-NI	NI-L	NG-HD	HD-NI	NI-L	HD-NG	NG-S	
TALE Control	NI-L	HD-NI	HD-NI	HD-HD	NI-L	HD-NI	NG-HD	NI-L	NG-NG	NI-L	NG-NG	NG-S	

Table S2.5 Target DNA sequences of TALEs in the endogenous promoter regions

TALE ID	Gene Name	Off-target/On-Target	Target DNA Sequence in Promoter
TALE-OSGIN2	OSGIN2	On-target	TCCTCCCCACCTTTAATTTT
TALE-ZC3H10	ZC3H10	On-target	TACCATATCCCATCCAACTC
TALE 5	OSGIN2	Off-target	TAAAATACCTGCTCATCACA
TALE 5	CRYBG3	Off-target	TAGCTTCCATTTTCAACACA
TALE 5	IL8	Off-target	TAAATTACCTCCCCAATAAA
TALE 5	Spats2l	Off-target	TAAATTATATTATCCACACA
TALE 5	PRC1	Off-target	TAACTTACCTATTCACCCCC
TALE 8	ZC3H10	Off-target	TACCATATCCCATCCAACTC
TALE 8	CAP2	Off-target	TAAAGTAACCAAACCCACTT
TALE 8	TMEM14C	Off-target	TATTATCTCCATTCCCACTT
TALE 8	VGLL4	Off-target	TATAATATCCATTTACACTT

Table S2.6 Reverse and Forward strand primer sequences used in qPCR experiments

RefSeq ID	Gene Name	Primer Sequences (5' to 3')	
		Forward Primer	**Reverse Primer**
NM_004337	OSGIN2	GCGCGAGGAAATGCCCAAAGAAAC	TCGCCCCAAGTTGTCACCAAAGT
NM_153605	CRYBG3	TTCGAGGTTGCTGGCTCCTCT	GCTGGGCAACCGCAGGAAGT
NM_000584	IL8	TGACTTCCAAGCTGGCCGTGG	ACTGCACCTTCACACAGAGCTGC
NM_001100422	Spats2l	ACCCGAGAGAGGCGTGAGCA	ATAGGCCCTGGGAATCCACAGCAA
NM_003981	PRC1	ACCTGGAGCTCAACGGCAGC	AGGGACGGATCCTTCGCAAACTC
NM_032786	ZC3H10	TCGGGTAGGCGGCTCTTTGT	ACATCGCTGCTGGGTTCTGCC
NM_006366	CAP2	AGCTGTCAGCCGCCTGGAGT	GGAGGGTGCCACACCTGCAAT
NM_016462	TMEM14C	CTGCGCAGGCACAACAGAGC	GCCTGCACCGGTCTCACGAA
NM_001128219	VGLL4	TGGGGCAAAAGCAAAGAGCTGGT	GCAGCTTCGCCTTCGTAGCACA
NM_002046	GAPDH	GAAATCCCATCACCATCTTCCAGG	GAGCCCCAGCCTTCTCCATG

Table S2.7 Constructs used in TALE repressor- shRNA co-expression experiments

Off-target shRNAs and TALEs are used to validate fold repression achieved for TALE only and shRNA only experiments, respectively. An off-target shRNA and TALE combination is used to determine the unrepressed CFP signal for each reporter.

Experiment/Reporter	BS-8 / FF4'	BS-8 / FF6'	BS-5 / FF4'	BS-5 / FF6'
Off-Target Control	TALE 5R + shRNA FF6	TALE 5R + shRNA FF4	TALE 8R + shRNA FF6	TALE 8R + shRNA FF4
shRNA only	TALE 5R + shRNA FF4	TALE 5R + shRNA FF6	TALE 8R + shRNA FF4	TALE 8R + shRNA FF6
TALE only	TALE 8R + shRNA FF6	TALE 8R + shRNA FF4	TALE 5R + shRNA FF6	TALE 5R + shRNA FF4
TALE+shRNA	TALE 8R + shRNA FF4	TALE 8R + shRNA FF6	TALE 5R + shRNA FF4	TALE 5R + shRNA FF6

Supplementary Methods

Analysis of potential TALE binding sites

To demonstrate the importance of considering off-target effects when designing synthetic TALEs to target endogenous genes, we analyzed the promoter regions of the human genome for the number of potential sites in each promoter region for which 12 bp, 18 bp and 20 bp TALE binding sites specific to a promoter can be designed. We found that TALEs designed to target 12 bp nucleotides sequences are unable to specifically target the promoter of any single gene as the target binding sites are too short, even if the promoter region is restricted to 200 bp from the TSS, and a designed TALE will always have an off-target binding site in the promoter of another gene. By extending the target nucleotide sequence to 18 bp it becomes possible, though difficult, to design TALEs specific to a promoter for a fraction of genes. If the promoter region is restricted to the 200 bp region upstream of the TSS, only 84% of genes have at least one target site in the

promoter region for which TALEs specific to that promoter can be designed. Of these 84% of genes, only 35% have more that 10 promoter specific 18 bp target sites, which is a small sequence space for attempting to design synthetic TALEs that function in a promoter specific manner. If the targetable promoter region for these TALEs is expanded to 500 bp, fewer than 7% of genes have more than 10 potential 18 bp TALE target sites that are specific (Figure S2.4). For the case where TALEs are presumed to function when they bind to an 18 bp sequence within the 2000 bp region upstream of the TSS, it is not possible to design a TALEs that interacts specifically with only one promoter. However, when we searched for potential 20 bp binding sites to which synthetic TALEs can be targeted without binding to off-target promoter sequences, we find that the design space is much larger. For 20 bp target sequences, 98% of human genes have more that 10 promoter specific binding sites present in 500 bp upstream regions, and many options are available to design TALEs that specifically target the expression of nearly every human gene (Figure S2.5). This suggests that future efforts seeking to utilize TALEs without unwanted off-target effects should consider designing proteins that bind target sites of at least 20 bp.

An algorithm to compute orthogonal 20-mers

We used Boolean algebra and symbolic modeling techniques to compute 20-mers at a given hamming distance from a set of genomic 20-mers. We translated N-mers, representing DNA sequences of length N over the alphabets {A, C, G, T}, into a Boolean vector **x** of length 2N. Boolean encoding 00, 01, 10 and 11 is used to represent alphabets A, C, G and T respectively. Using this Boolean encoding, a base pair at position *i* in a given N-mer is represented by two consecutive bit positions x_{2i-1} and x_{2i} of the

corresponding Boolean vector $\mathbf{x}$. In order to perform efficient set operations such as intersection, a set of N-mers can be represented by Boolean functions. An example demonstrating a Boolean representation of N-mers and a Boolean function representation of a set of N-mers is given below.

Example: Given a set of 2-mers, S = {TT, AA}, its Boolean representation is given by S^B= $\{x_1x_2x_3x_4 \ , \bar{x}_1\bar{x}_2\ \bar{x}_3\bar{x}_4\}$, where $\bar{x}_1$ stands for logic negation of x_1. The corresponding Boolean function for set S^B is given by:

$$F^B = [(x]_1 \wedge x_2 \wedge x_3 \wedge x_4) \vee [(\bar{x}]_1 \wedge \bar{x}_2 \wedge \bar{x}_3 \wedge \bar{x}_4)$$

The hamming distance between two N-mers is defined as the number of base pair positions at which the two N-mers differ. The hamming distance for Boolean vector encoding of N-mers is defined as the number of consecutive bit positions at which the two Boolean vectors are different. For example, the hamming distance between Boolean vectors 0001 and 0110 is 2 as the two vectors differ at positions {1,2} and {3,4}. The hamming distance between two Boolean vectors $\mathbf{x}$ and $\tilde{\mathbf{x}}$ of length N can be calculated using the Boolean function in Equation S-I, where symbols Θ and $\vee$ stand for Boolean operations XOR and OR respectively.

$$H(\mathbf{x}, \tilde{\mathbf{x}}) = \sum_{i=1}^{N} [[(\mathbf{x}]_{2i-1} \oplus \tilde{\mathbf{x}}_{2i-1}) \vee [(\mathbf{x}]_{2i} \oplus \tilde{\mathbf{x}}_{2i})] \qquad \text{(S-I)}$$

Our algorithm first computes the Hamming distance function between Boolean vectors corresponding to all possible genomic N-mers. Then, given a Boolean vector $\mathbf{x}$ and the hamming distance function H, all Boolean vectors that are above a given hamming distance d from the vector $\mathbf{x}$ can be quickly computed by restricting the

hamming distance function H to threshold d (represented by H^d) and applying the Boolean function in Equation S-II.

$$F_d^B = \underset{x_i \in \mathbf{x}}{\exists} \left[H^d(\mathbf{x}, \tilde{\mathbf{x}}) \wedge \mathbf{x} \right] \tag{S-II}$$

The symbol ∃ in Equation S-II represents existential quantification, which essentially drops all variables in $\mathbf{x}$ by iteratively assigning them values of 0 and 1. The resulting Boolean function F_d^B, after applying the existential quantification, represents all Boolean vectors at the hamming distance above the threshold d from the input vector $\mathbf{x}$. In order to compute N-mers at a given hamming distance from every element in a set of Boolean vectors, we apply Equation S-II to all Boolean vectors in the set and then take the intersection of resulting Boolean functions.

In order to efficiently represent a large set of Boolean vectors and perform Boolean operations on them, we use symbolic representation and modeling techniques based on Reduced Ordered Binary Decision Diagrams (ROBDDs) (46,47). ROBDDs are directed top-to-bottom graphs, where a top node represents the Boolean function being evaluated, each intermediate node represents a Boolean variable, and two leaf nodes (0 and 1) represent whether the Boolean function evaluates to true (i.e. 1) or false (i.e. 0). Every intermediate node has two outgoing edges representing whether the Boolean variable is assigned the value ‘0’ or ‘1’. Each path from the top node to the leaf node has the same order of Boolean variables and represents a Boolean vector assignment. A path leading to leaf node ‘1’ or ‘0’ represents the Boolean variable assignment for which Boolean function evaluates to true or false respectively. An example of a ROBDD representing the set of Boolean vectors in the example given above is shown in Figure

S2.6. All Boolean logic functions such as AND, OR and NOT can be efficiently performed on ROBDD representations of Boolean functions.

The hamming distance function in Equation S-I can also be represented using Algebraic Decision Diagrams (ADDs), which are a modified form of ROBDDs such that the leaf nodes can take values from the set of real numbers instead of only Boolean values 0 and 1 (46). In our case, ADDs corresponding to the hamming distance function in Equation S-I will take integer values from 0 to N representing the number of base pairs at which the two N-mers differ. An ADD representing the hamming distance between the set of N-mers in the above example is shown in Figure S2.7.

Our algorithm first constructs an ADD to represent the hamming distance between Boolean vectors of all possible N-mers. In our case N=20, therefore such an ADD would represent the pair-wise hamming distance between all possible 4^{20} 20-mers. If one is interested in only Boolean vectors that have a hamming distance above a threshold *d*, then an ADD can be converted to a corresponding ROBDD by changing all the leaf nodes above the value *d* to '1' and all the remaining leaf nodes to '0'. The widely used Binary Decision Diagrams modeling package CUDD, was used to perform all ROBDDs and ADDs representation and manipulation in this manuscript (48).

Cloning scheme for TALEs

Restriction sites for EcoRI, XbaI, PstI and Alw26I were designed at the 5'end and restriction sites for BbsI, SpeI, NotI and PstI were designed at the 3'end in all sub-modules corresponding to repeat domains. The constant 5'- region and constant 3'-region sub-modules had similar restriction sites to repeat domains sub-modules except that Alw26I and BbsI sites are absent in the constant 5'- and 3'- region constructs,

respectively. The pidSmart vector (Integrated DNA Technology, Coralville, IA) was used for cloning and assembling TALE fragments into complete TALE sequences using standard molecular biology techniques and all DNA constructs were verified by sequencing. Figure S2.2 further describes the TALE assembly scheme.

Appendix III

Protein Scaffold-Activated Protein Trans-Splicing in Mammalian Cells

Daniel F. Selgrade,[1] Jason J. Lohmueller[1,2], Florian Leinert[1,2] & Pamela A. Silver[1,2]

[1] *Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA*

[2] *Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts 02115, USA*

Reprinted (adapted) from Selgrade DF, Lohmueller JJ, Lienert F, Silver PA. (2013). Protein Scaffold-Activated Protein Trans-Splicing in Mammalian Cells. *J Am Chem Soc.* May 8. [Epub ahead of print].

Jason J. Lohmueller designed all experiments and wrote the manuscript with input from Daniel F. Selgrade. Daniel F. Selgrade carried out all experiments and produced all of data with mentorship from Jason J. Lohmueller and collaboration with Florian Leinert.

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

pubs.acs.org/JACS

Protein Scaffold-Activated Protein Trans-Splicing in Mammalian Cells

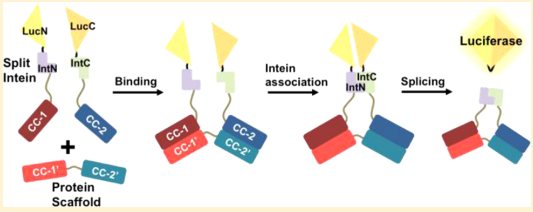
Daniel F. Selgrade,[†,§] Jason J. Lohmueller,[†,‡,§] Florian Lienert,[†,‡] and Pamela A. Silver*[,†,‡]

[†]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, United States

[‡]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts 02115, United States

Supporting Information

ABSTRACT: Conditional protein splicing is a powerful biotechnological tool that can be used to rapidly and post-translationally control the activity of a given protein. Here we demonstrate a novel conditional splicing system in which a genetically encoded protein scaffold induces the splicing and activation of an enzyme in mammalian cells. In this system the protein scaffold binds to two inactive split intein/enzyme extein protein fragments leading to intein fragment complementation, splicing, and activation of the firefly luciferase enzyme. We first demonstrate the ability of antiparallel coiled-coils (CCs) to mediate splicing between two intein fragments, effectively creating two new split inteins. We then generate and test two versions of the scaffold-induced splicing system using two pairs of CCs. Finally, we optimize the linker lengths of the proteins in the system and demonstrate 13-fold activation of luciferase by the scaffold compared to the activity of negative controls. Our protein scaffold-triggered conditional splicing system is an effective strategy to control enzyme activity using a protein input, enabling enhanced genetic control over protein splicing and the potential creation of splicing-based protein sensors and autoregulatory systems.



INTRODUCTION

Protein splicing is a post-translational modification that can control the activity of a protein by assembling it from inactive fragments. Analogous to RNA splicing, protein splicing is the process by which an intervening protein domain, or intein, self-excises out of a larger polypeptide, ligating the two flanking polypeptides—termed exteins—into a single protein.[1] Protein splicing can occur in *cis*- or in *trans*-.[2−4] For trans-splicing the intein sequence is split into two fragments, and the splicing reaction occurs between two distinct polypeptides. Intein domains are highly modular and have been used by researchers to assemble several different proteins in various experimental contexts.[5−8]

The development of conditional protein splicing has enabled researchers to post-translationally control protein activity in response to specific molecular inputs and has already become a useful research tool. By engineering split inteins that contained the rapamycin ligand-binding domains, Mootz et al. first demonstrated that trans-splicing could be induced by the small molecule rapamycin.[6] Other conditional splicing systems have been generated to induce splicing in response to temperature, light, and chemical ligands such as 4-HT.[9−13]

Here we propose a novel conditional splicing system in which splicing is induced by the presence of a genetically encoded protein scaffold. In contrast to previously developed conditional splicing systems activated by exogenously administered inducers (small molecules and light), a protein inducer can be directly linked to endogenous biological pathways. This characteristic enables the potential to monitor or rewire biological pathways at the protein level. The engineering of synthetic post-translational signaling pathways reflects a common strategy used by natural systems and is a major research focus. However, only a few general approaches exist.[14−18] The protein-induced protein splicing system offers a general mechanism to post-translationally convert a protein input into a fully formed output protein through engineered binding domains. The general reaction mechanism demonstrated by our work also suggests methods to create splicing-based protein sensors and autoregulatory systems.

Biomolecular scaffolds have previously been used to increase the yield of biochemical synthesis pathways by bringing together enzymes operating in a pathway.[19−22] We hypothesized that protein scaffolds could also be used to bring together split inteins and trigger protein splicing. Our system consists of two fusion proteins—each containing split intein/extein domains fused to a scaffold binding domain—and the input scaffold protein. In the presence of the scaffold the two fusion proteins bind to the scaffold leading to association of the intein fragments, splicing, and activation of the output protein.

We constructed, tested, and optimized a scaffold-induced splicing system comprising well-characterized protein components. We used two pairs of previously characterized antiparallel coiled-coils termed LZA/LZB and EE/RR.[23−25] LZA is known to bind strongly to LZB and EE to RR, but no binding is expected to occur between proteins in the different pairs. These coiled-coils drive the specific association of the intein/extein fusion proteins and the synthetic scaffold. We

Received: February 15, 2013

   dx.doi.org/10.1021/ja401689b | J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

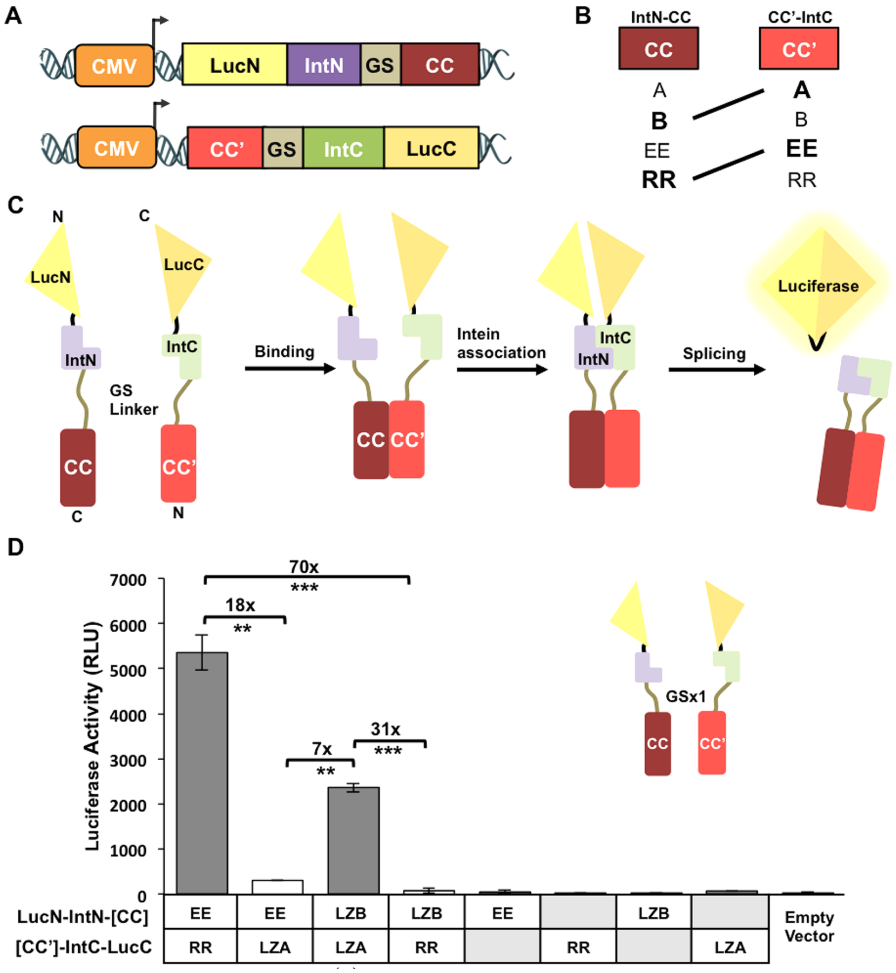


Figure 1. Coiled-coil-mediated protein trans-splicing. (A) Schematic representation of the expression system used to test trans-splicing of CC-intein/extein fusion proteins. Each CC-intein/extein fusion protein is expressed from the CMV promoter. Coiled-coil domains CC and CC′ are fused to split inteins and luciferase exteins, IntN/LucN and LucC/IntC, via flexible glycine-serine linkers (GS). (B) Schematic representation of the complementary CC binding assay design. CC-intein/extein pairs tested are in boldface, and those connected by lines are expected to interact. (C) Binding of complementary coiled-coil domains leads to intein fragment complementation and splicing and activation of firefly luciferase. (D) Activity of recombinant intein pairs as measured by luciferase output. Shaded boxes represent transfection with an empty expression vector. Data are presented as mean ± s.d., $n = 3$. Two asterisks, $P < 0.01$; three asterisks, $P < 0.001$.

chose to use the *Saccharomyces cerevisiae* vacuolar ATPase (VMA) split intein because these split fragments display very weak splicing activity in the absence of outside protein-binding domains.[8,26] For the output protein we used firefly luciferase, as it has a sensitive biochemical readout and previously determined extein split sites.[8,27] After demonstrating the ability of the CCs to mediate splicing, we optimized the linker lengths of the individual component proteins of the system. Finally, we showed that our protein scaffold system had comparable efficacy to that of the established rapamycin-induced splicing system.[8]

RESULTS AND DISCUSSION

We first sought to determine whether coiled-coil binding could mediate trans-splicing between proteins containing split VMA intein fragments in mammalian cells. We generated *Cytomegalovirus* (CMV) promoter driven expression plasmids encoding fusion proteins comprising a CC domain fused by a flexible glycine-serine (GS) linker to an N- or C-terminal split VMA intein domain and an N- or C-terminal firefly luciferase extein fragment, respectively (Figure 1A). Amino acid sequences of the coiled-coil domains are listed in Supporting Information, Table S1. We expected that complementary CCs (LZA with LZB and EE with RR) would bind, leading to split intein complementation, protein splicing, and luciferase activity. Conversely, we anticipated that proteins with noncomplementary CCs would not interact, resulting in inefficient splicing and low luciferase activity (Figure 1B,C).

To test for CC-mediated splicing we transiently cotransfected different combinations of N-intein and C-intein expression plasmids into U2OS osteosarcoma cells and assayed

dx.doi.org/10.1021/ja401689b | J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

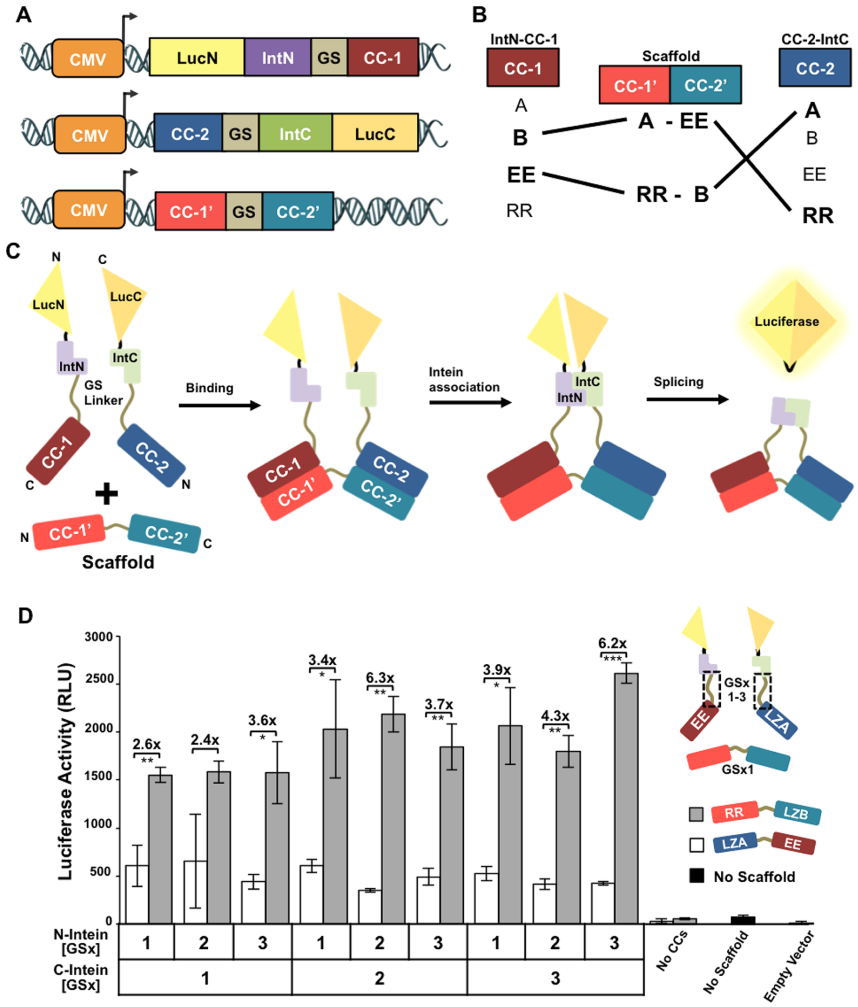


Figure 2. Protein scaffold-activated protein trans-splicing. (A) Schematic representation of the expression system used to test conditional splicing of CC-intein/extein fusions in response to CC scaffolds. (B) Schematic representation of the expected complementary CC binding for each synthetic scaffold and CC-intein/extein pair. Proteins tested are in boldface, and those connected by lines are expected to interact. (C) Recombinant inteins containing noncomplementary coiled-coil domains CC-1 and CC-2 associate only in complex with a protein scaffold containing complementary CCs, CC-1′ and CC2′. Presence of the scaffold results in the splicing and activation of firefly luciferase. (D) Induction of splicing between EE and LZA split inteins of varying GS linker lengths by a protein scaffold. Constructs encoding each intein pair were cotransfected with a construct encoding an ON-target scaffold RR-GS1-LZB (gray) or an OFF-target scaffold LZA-GS1-EE (white). The 'No Scaffold' control consists of the 3 GS linker EE and LZA split inteins transfected with an empty vector instead of a scaffold. The 'No CCs' control represents cotransfections of inteins not fused to coiled-coils. Data presented as mean ± s.d., $n = 3$. One asterisk, $P < 0.05$; two asterisks, $P < 0.01$; three asterisks, $P < 0.001$.

for luciferase activity at 48 h. The specific CC combinations that we built and transfected are shown in boldface, and connecting lines indicate expected interactions in Figure 1B.

We found that luciferase activity was significantly higher when the coexpressed proteins contained complementary CCs, indicating that CCs can coordinate specific protein trans-splicing. Proteins containing the EE/RR pair showed higher overall luciferase activity than the LZA/LZB pair (5360 RLU and 2360 RLU, respectively) (Figure 1D). This result agrees with higher activities reported for other protein systems using the EE/RR CC pair compared to those using the LZA/LZB pair.[23,26]

Importantly, Schwartz et al. previously demonstrated that any luciferase activity observed from the intein/extein pairs used in the scaffold system is due to protein splicing and not protein fragment complementation.[8] Expression constructs encoding intein/extein protein fragments lacking CCs and single CC-intein/extein fragment transfections showed no significant luciferase activity above the vector-only control, further supporting the role of coiled-coil binding in mediating the splicing reaction (Figure 1D). Given the large number of CC pairs present in the literature, these results suggest that combining CCs with the VMA split intein fragments could be a general strategy to produce a large number of new functionally

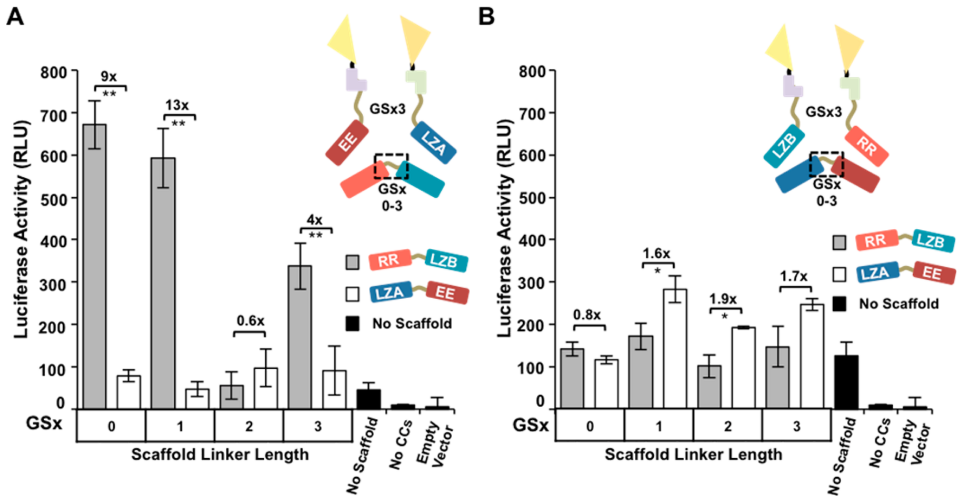


Figure 3. Effects of scaffold linker length on scaffold-induced protein splicing. (A) Luciferase activity of the LZA-GS3-intC/LucC and LucN/IntN-GS3-EE proteins induced by CC scaffolds with GS linkers of varying lengths. CCs for both ON-target (RR-LZB) and OFF-target (EE-LZA) scaffolds were fused together by 0−3 GS linkers. (B) Luciferase activity of the RR-GS3-LucC/LucC and LucN/IntN-GS3-LZB proteins induced by CC scaffolds with GS linkers of varying lengths. CCs for both ON-target (EE-LZA) and OFF-target (RR-LZB) scaffolds were fused together by 0−3 GS linkers. The 'No Scaffold' control consists of the intein pair transfected with an empty vector instead of a scaffold. The 'No CCs' control represents cotransfections of inteins not fused to coiled-coils. Data presented as mean ± s.d., $n = 3$. One asterisk, $P < 0.05$; two asterisks, $P < 0.01$; three asterisks, $P < 0.001$.

orthogonal intein pairs with minimal cross-reactivity.[28] Additionally, the low levels of luciferase activity for protein pairs with noncomplementary CCs indicate that these proteins could potentially serve as substrates in a scaffold-induced conditional splicing system.

Next, we investigated whether we could induce trans-splicing of two noninteracting CC-intein/extein proteins using a protein scaffold. In this system the scaffold protein is comprised of two CCs that are complementary to the CCs of the CC-intein/extein fragments. We expect both CC-intein/extein fragments to bind to the scaffold leading to split intein complementation, splicing of the luciferase exteins, and luciferase activity (Figure 2A,B,C). We first focused on scaffolding the LZA/EE pair of CC-intein/extein fragments. As previous results indicated the importance of spacing between scaffolded molecules, we created scaffolds and CC-intein/extein proteins with flexible linkers of varying lengths.[19,20,29] We generated CMV expression constructs encoding fusion proteins with 1-, 2-, or 3-copies of the glycine-serine linker $(GGGS)_3$ between CCs and the split intein/extein domains (Figure 2A). To test for scaffold-mediated splicing and the effect of CC-intein/extein linker length we cotransfected different combinations of expression plasmids encoding the N-intein and C-intein proteins with different linker lengths along with an 'ON-target' RR-LZB scaffold or an 'OFF-target' LZA-EE scaffold. We assayed for luciferase activity at 48 h. The expected CC interactions for the scaffolds and intein fragments that we tested (in boldface) are shown with connecting lines in Figure 2B.

The results of the luciferase assays demonstrated that the protein scaffold could induce specific trans-splicing of two noninteracting intein/extein proteins. We found that the ON-target scaffold led to significantly higher levels of luciferase activity than the OFF-target scaffold for all linker lengths of the LZA/EE proteins tested (2.4−6.3 fold). The CC-intein/extein proteins with 3-GS linkers exhibited the highest luciferase levels (~2600 RLU) (Figure 2D). These 3-GS linker proteins also exhibited significantly higher levels of luciferase in the presence of the ON-target scaffold compared to the 'No Scaffold' control. Neither scaffold affected the splicing and luciferase activity of the control intein/extein proteins containing no CCs. We also found that the behavior of this system was robust to changes in amounts of DNA transfected (Supporting Information [SI], Figure 1).

Next, we investigated the effect of the scaffold linker length on scaffold-induced splicing. We created DNA constructs encoding scaffolds with 0−3 × GS linkers. We also generated expression constructs encoding CC-GS3-intein/extein proteins with the LZB/RR CC pair. The full list of constructs generated and their subparts are listed in SI, Table S2. We transfected these CC-intein/extein constructs and scaffolds with different linker lengths and assayed for luciferase activity.

We found that most of the scaffolds of different GS-linker lengths were capable of inducing splicing and that the linker length had variable effects on luciferase activity (Figure 3A). For the LZA/EE-intein/extein system the luciferase activity was highest for the shortest, 0-GS linker, scaffold and correlated negatively with the length of the scaffold GS linker. The 0-, 1-, and 3-GS scaffolds all induced higher splicing levels than both the OFF-target scaffold and 'No Scaffold' controls. The 2-GS scaffold showed no significant induction of splicing activity, possibly due to steric constraints (Figure 3A).[29] While the induction levels for the LZB/RR-intein/extein system were lower than those of the LZA/EE system, the 1-and 2-GS LZB/RR systems showed significantly higher levels of luciferase with ON-target scaffolds compared to 'OFF-target' and 'No Scaffold' controls (Figure 3B). Of note, observed differences in the activity of the LZA/EE-intein/extein system in Figures 2D, 3A, 4B are most likely due differences in the luciferase assay kits used. While experiments within a single figure used the same

dx.doi.org/10.1021/ja401689b | J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

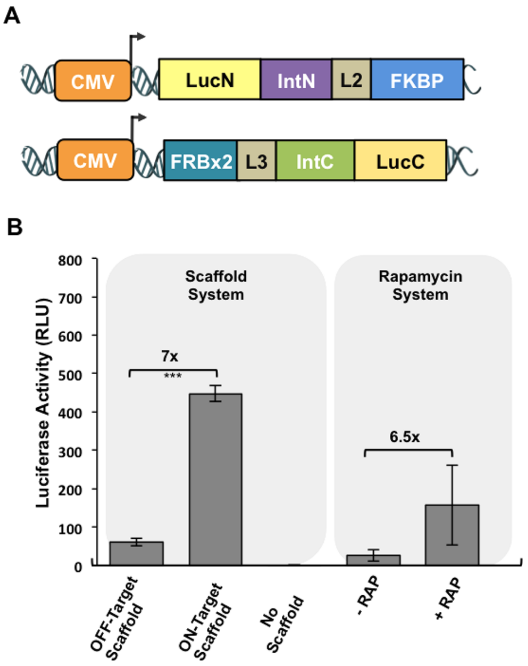


Figure 4. Comparison of the protein scaffold-induced splicing system to the rapamycin-induced splicing system. (A) Schematic representation of the expression system used to test rapamycin-induced conditional splicing. Rapamycin-binding domains FRB (two copies) and FKBP were fused via flexible linkers L2 and L3 to split intein-firefly luciferase extein fusions.[8] (B) Comparison of luciferase activity of the rapamycin-induced splicing system and the scaffold-induced system (LZA-GS3-LucC and LucN-GS3-EE and GS1 scaffolds). The 'No Scaffold' control consists of the intein pair transfected with an empty vector instead of a scaffold. Data presented as mean ± s.d., n = 3. One asterisk, $P < 0.05$; two asterisks, $P < 0.01$; three asterisks, $P < 0.001$.

assay kit and can be quantitatively compared, results between figures should not be directly compared.

Finally, we compared the efficacy of the scaffold-induced splicing system to that of the established rapamycin-induced splicing system reported in Schwartz et al. In this system the rapamycin-binding domains FRB and FKBP bind simultaneously to rapamycin, leading to split intein complementation, protein splicing, and luciferase activity. We cloned CMV expression plasmids encoding the rapamycin inducible system, transfected it into U2OS cells and assayed for activity following 48 h. For the rapamycin system transfected cells were incubated with rapamycin or DMSO-only (Figure 4A). We found that rapamycin successfully induced luciferase activity 6.5-fold compared to the DMSO-only vehicle control. In comparison, under the same experimental conditions our scaffold-induced system exhibited similar behavior with a 7-fold induction in the presence of scaffold (Figure 4B). It should be noted that the rapamycin-inducible system has been previously shown to have optimal activity at 25 °C in *Drosophila* S2 cells; however, whenever implemented in mammalian cells, reactions are performed at 37 °C.[8] These results demonstrate that the efficacy of our scaffold-induced system is similar to that of an established conditional splicing system.

The successful implementation of our protein scaffold-induced splicing system provides the mechanistic foundation for further adaptations and applications. As the VMA split inteins have been used in various organisms and contexts, it is likely that the scaffold-induced system could also be adapted to control protein activity in a number of instances.[5−8] Additionally, the modularity of the VMA split intein with respect to binding domains and extein proteins suggests that other CCs or protein binding domains such as the PDZ, SH3, or zinc finger domains could be used and that the system could be adapted to splice together other output proteins.[30−33] That the scaffold is a protein allows for the potential to link its presence—and thus also the activation of the splicing reaction—to other biological processes *in vivo*. This goal could be accomplished by expressing the scaffold from a promoter specific to the desired process, or by post-translationally controlling its activity or localization.[14] Converting the presence of one protein into the production of a second protein also implies a design for a general protein sensor in which the protein being sensed serves as the scaffold protein for the splicing reaction (Figure 5A). Finally, the efficacy of this protein-induced system using two separate pairs of inteins with little cross-reactivity provides the basis for higher-order functioning systems. These systems could include autoregulatory networks such as protein splicing cascades or amplifiers analogous to established DNA-based systems (Figure 5B).[34]

■ CONCLUSIONS

We have developed a novel conditional splicing system in which a genetically encoded protein scaffold induces the trans-splicing and activation of an enzyme. We first demonstrated the ability of complementary CCs to mediate protein trans-splicing between two proteins, effectively generating two new sets of orthogonal inteins. We then demonstrated the efficacy of the scaffold-induced splicing system with two different sets of complementary CC proteins. We optimized the system based on component linker lengths, yielding a system capable of 13-fold induction over the 'No scaffold' and 'OFF-target' controls. Finally, we showed that the system had comparable efficacy to that of an established conditional protein splicing system. In sum, the protein scaffold-induced splicing system adds to the repertoire of modular approaches that researchers can employ to precisely control protein activity and biological functions in living cells.

■ EXPERIMENTAL SECTION

Recombinant DNA constructs. Recombinant plasmids were created using the Biobrick Cloning method. DNA sequences encoding the VMA intein fragments, LZA, LZB, EE, RR, FRB, FKBP were flanked with Biobrick ends and synthesized for order by Integrated DNA Technologies (IDT). The intein−luciferase fusion parts were constructed via PCR and BspQI restriction enzyme methods. The complete list of constructs and their constituent BioBrick parts can be found in Table S2 of SI. Sequences of all BioBrick subparts are listed in Table S3 of SI. For CMV expression constructs coding regions were cut with Xba1 and Not1 and cloned into the NheI and NotI sites of the CMV expression plasmid "pCDNA5ins."[35]

Mammalian Cell Culture and Transfection. Human osteosarcoma-derived U2OS cells (ATCC no. HTB-96) were cultured at 37 °C, 5% CO_2 in growth medium (McCoy's 5A medium, 10% FBS, 100 U/mL penicillin, and 100 mg/mL streptomycin). For transfections, cells were plated in 12-well plates at ∼150,000 cells per well in 1 mL growth medium. Transient transfections were performed 24 h after plating at 80% confluency using Lipofectamine LTX with Plus

dx.doi.org/10.1021/ja401689b | J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

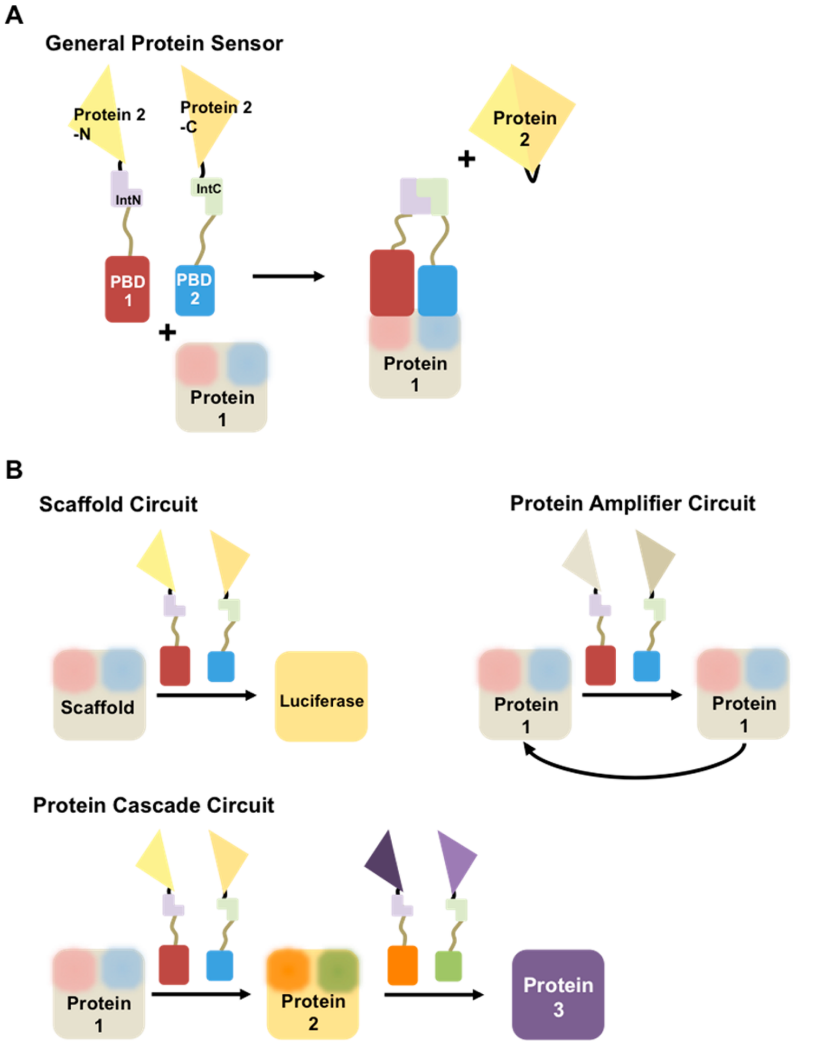


Figure 5. Potential applications of the protein-induced protein splicing system. (A) Schematic representation of a general protein sensor. The sensor comprises two fusion proteins consisting of protein binding domains (PBDs) for an arbitrary protein - Protein 1, split intein fragments IntN and IntC, and exteins of the output protein - Protein 2. Upon binding to Protein 1, the intein fragments complement and Protein 2 is spliced and activated. (B) Potential protein-based autoregulatory networks. The scaffold circuit described in this manuscript is represented in abbreviated diagrammatic form. In a hypothetical protein amplification circuit, Protein 1 induces the splicing of additional Protein 1, creating positive feedback and amplification of Protein 1. For a protein cascade circuit, Protein 1 induces the splicing of Protein 2 which in turn catalyzes the splicing of Protein 3.

(Invitrogen) and a total of 1 μg DNA per reaction according to the manufacturer's protocol. All transfections were performed in triplicate with the precise DNAs and amounts as specified in Table S4 of SI. Transfected cells were incubated at 37 °C for 48 h prior to analysis by luciferase assay.

Luciferase Assay. Luciferase activity of transfected cells was measured using Dual Luciferase Reporter Assay (Promega) according to manufacturer's instructions. Briefly, a luciferase lysis buffer (1 × passive lysis buffer supplemented with 1 μM ZnCl) was used to lyse the cells and inhibit background splicing. To each transfection well we added 250 μL of the modified buffer, and the plates sat on the shaker for 15 min prior to aliquotting in a 96-well plate. For each transfection well, 100 μL of Luciferase Assay Buffer (LARII) was pipetted over 20 μL of cell lysate and photometer readings were taken for each well. Luciferase activity is reported in relative light units (RLU) calculated by subtracting the raw output of each transfection well by an initial blank value taken on a well containing only cell lysate. All charts contain data collected in a single assay run using pooled LARII buffer.

■ ASSOCIATED CONTENT

Supporting Information

DNA titration of LZA-EE scaffold system, DNA constructs and subparts, DNA transfection amounts, DNA subpart sequences, amino acid sequences of the coiled-coil domains. This material is available free of charge via the Internet at http://pubs.acs.org.

dx.doi.org/10.1021/ja401689b | J. Am. Chem. Soc. XXXX, XXX, XXX–XXX

AUTHOR INFORMATION

Corresponding Author

Pamela_Silver@hms.harvard.edu

Author Contributions

[§]D.F.S. and J.J.L. contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by funds from the NSF SYNBERC (J.J.L.), The Harvard College Research Program (D.F.S), EMBO and Human Frontier Science Program fellowship (F.L.), and the Wyss Institute for Biologically Inspired Research (P.A.S.). We acknowledge all members of the Silver lab for helpful comments and discussion. We also acknowledge J. Torella and D. B. Thompson for carefully reading the manuscript.

REFERENCES

(1) Paulus, H. *Annu. Rev. Biochem.* **2000**, *69*, 447.
(2) Perler, F. B.; Davis, E. O.; Dean, G. E.; Gimble, F. S.; Jack, W. E.; Neff, N.; Noren, C. J.; Thorner, J.; Belfort, M. *Nucleic Acids Res.* **1994**, *22*, 1125.
(3) Wu, H.; Xu, M. Q.; Liu, X. Q. *Biochim. Biophys. Acta* **1998**, *1387*, 422.
(4) Sun, W.; Yang, J.; Liu, X. Q. *J. Biol. Chem.* **2004**, *279*, 35281.
(5) Buskirk, A. R.; Ong, Y. C.; Gartner, Z. J.; Liu, D. R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 10505.
(6) Mootz, H. D.; Blum, E. S.; Tyszkiewicz, A. B.; Muir, T. W. *J. Am. Chem. Soc.* **2003**, *125*, 10561.
(7) Mootz, H. D.; Blum, E. S.; Muir, T. W. *Angew. Chem., Int. Ed.* **2004**, *43*, 5189.
(8) Schwartz, E. C.; Saez, L.; Young, M. W.; Muir, T. W. *Nat. Chem. Biol.* **2007**, *3*, 50.
(9) Zeidler, M. P.; Tan, C.; Bellaiche, Y.; Cherry, S.; Hader, S.; Gayko, U.; Perrimon, N. *Nat. Biotechnol.* **2004**, *22*, 871.
(10) Berrade, L.; Kwon, Y.; Camarero, J. A. *ChemBioChem* **2010**, *11*, 1368.
(11) Skretas, G.; Wood, D. W. *Protein Sci.* **2005**, *14*, 523.
(12) Peck, S. H.; Chen, I.; Liu, D. R. *Chem. Biol.* **2011**, *18*, 619.
(13) Adam, E.; Perler, F. B. *J. Mol. Microbiol. Biotechnol.* **2002**, *4*, 479.
(14) Bashor, C. J.; Helman, N. C.; Yan, S.; Lim, W. A. *Science* **2008**, *319*, 1539.
(15) Dueber, J. E.; Mirsky, E. A.; Lim, W. A. *Nat. Biotechnol.* **2007**, *25*, 660.
(16) Wehr, M. C.; Laage, R.; Bolz, U.; Fischer, T. M.; Grunewald, S.; Scheek, S.; Bach, A.; Nave, K. A.; Rossner, M. J. *Nat. Methods* **2006**, *3*, 985.
(17) O'Shaughnessy, E. C.; Palani, S.; Collins, J. J.; Sarkar, C. A. *Cell* **2011**, *144*, 119.
(18) Grunberg, R.; Serrano, L. *Nucleic Acids Res.* **2010**, *38*, 2663.
(19) Delebecque, C. J.; Lindner, A. B.; Silver, P. A.; Aldaye, F. A. *Science* **2011**, *333*, 470.
(20) Dueber, J. E.; Wu, G. C.; Malmirchegini, G. R.; Moon, T. S.; Petzold, C. J.; Ullal, A. V.; Prather, K. L.; Keasling, J. D. *Nat. Biotechnol.* **2009**, *27*, 753.
(21) Ryadnov, M. G.; Woolfson, D. N. *J. Am. Chem. Soc.* **2004**, *126*, 7454.
(22) Bromley, E. H.; Sessions, R. B.; Thomson, A. R.; Woolfson, D. N. *J. Am. Chem. Soc.* **2009**, *131*, 928.
(23) Oakley, M. G.; Kim, P. S. *Biochemistry* **1998**, *37*, 12603.
(24) Moll, J. R.; Ruvinov, S. B.; Pastan, I.; Vinson, C. *Protein Sci.* **2001**, *10*, 649.
(25) Shekhawat, S. S.; Porter, J. R.; Sriprasad, A.; Ghosh, I. *J. Am. Chem. Soc.* **2009**, *131*, 15284.
(26) Chong, S.; Shao, Y.; Paulus, H.; Benner, J.; Perler, F. B.; Xu, M. Q. *J. Biol. Chem.* **1996**, *271*, 22159.
(27) Luker, G. D.; Sharma, V.; Pica, C. M.; Dahlheimer, J. L.; Li, W.; Ochesky, J.; Ryan, C. E.; Piwnica-Worms, H.; Piwnica-Worms, D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6961.
(28) Thompson, K. E.; Bashor, C. J.; Lim, W. A.; Keating, A. E. *ACS Synth. Biol.* **2012**, *1*, 118.
(29) Boyle, A. L.; Bromley, E. H.; Bartlett, G. J.; Sessions, R. B.; Sharp, T. H.; Williams, C. L.; Curmi, P. M.; Forde, N. R.; Linke, H.; Woolfson, D. N. *J. Am. Chem. Soc.* **2012**, *134*, 15457.
(30) Sheng, M.; Sala, C. *Annu. Rev. Neurosci.* **2001**, *24*, 1.
(31) Yu, H.; Chen, J. K.; Feng, S.; Dalgarno, D. C.; Brauer, A. W.; Schreiber, S. L. *Cell* **1994**, *76*, 933.
(32) Giesecke, A. V.; Fang, R.; Joung, J. K. *Mol. Syst. Biol.* **2006**, *2*, 2006.
(33) Armstrong, C. T.; Boyle, A. L.; Bromley, E. H.; Mahmoud, Z. N.; Smith, L.; Thomson, A. R.; Woolfson, D. N. *Faraday Discuss.* **2009**, *143*, 305.
(34) Yin, P.; Choi, H. M.; Calvert, C. R.; Pierce, N. A. *Nature* **2008**, *451*, 318.
(35) Lohmueller, J. J.; Armel, T. Z.; Silver, P. A. *Nucleic Acids Res.* **2012**, *40*, 5180.

Supplementary Information

Protein scaffold-activated protein trans-splicing in mammalian cells

Daniel F Selgrade[1,‡] Jason J Lohmueller[1,2,‡], Florian Lienert[1,2] & Pamela A Silver[1,2,*]

[1] Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[2] Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA

[‡] These authors contributed equally to this work.

* To whom correspondence should be addressed.
Email: pamela_silver@hms.harvard.edu

Table of Contents:

Supplementary Figure S1. The protein scaffold-induced system is robust to DNA concentration. DNA constructs encoding LZA-GS3-intC/LucC and LucN/IntN-GS3-EE were co-transfected in mammalian cells at five different concentrations with DNA encoding ON- or OFF- Target protein scaffolds. Both inteins were transfected at the specified concentrations while the scaffolds DNA concentration was held constant. 'No LZs' denotes 1x each of LucN/IntN and intC/LucC transfected with 1x of the specified synthetic scaffold. No scaffold represents LZA-GS3-IntC and IntN-GS3-EE transfected at 3x each. One asterisk, $P < 0.05$; two asterisks, $P < 0.01$; three asterisks, $P < 0.001$. Data presented as mean +/- s.d, n=3.

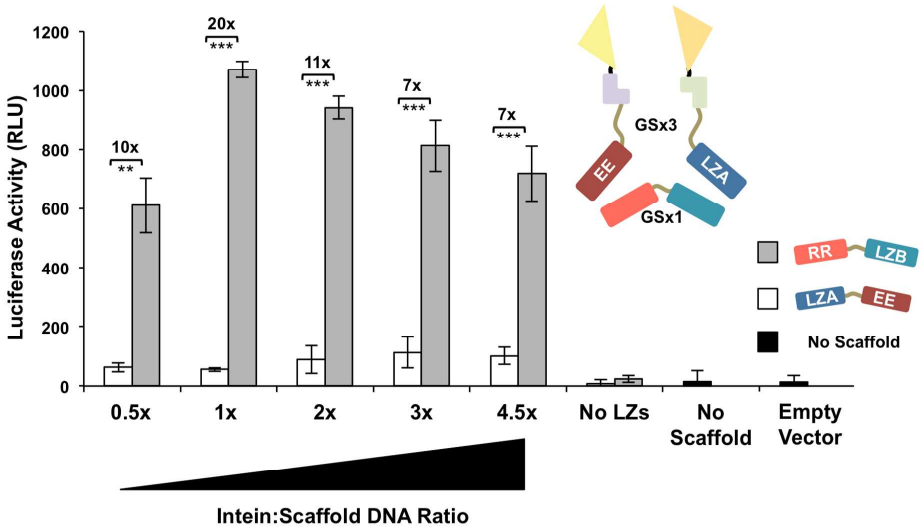


Table S1. Amino acid sequences of coiled coils.

Coiled-Coil	Amino acid Sequence
A	AQLEKELQALEKKLAQLEWE NQALEKELAQ
B	AQLKKKLQANKKELAQLKWK LQALKKKLAQ
EE	LEIEAAFLEQENTALETEVA ELEQEVQRLENIVSQYETRY GPL
RR	LEIRAAFLRRRNTALRTRVA ELRQRVQRLRNIVSQYETRY GPL

Table S2. Experimental DNA Constructs and their constituent subparts. DNA constructs are listed along with their constituent BioBrick subparts. Subparts were combined using BioBrick cloning and inserted into a mammalian expression vector as described in the *Experimental Section.* See *Table-Supp2.xls.*

Table S3. Sequences of DNA BioBrick subparts. BioBrick subparts used to make experimental DNA constructs. See *Table-Supp3.xls*.

Table S4. DNA plasmids co-transfected for each experiment. Plasmids and plasmid amounts transfected for each intein experiment. See *Table-Supp4.xls*.